AD_____

GRANT NUMBER: DAMD17-94-J-4332

TITLE: Statistical Methods for Analyzing Time-Dependent Events in Breast Cancer Chemoprevention Studies

PRINCIPAL INVESTIGATOR: Dr. George Y. C. Wong

CONTRACTING ORGANIZATION: Strang Cancer Prevention Center
New York, New York 10021

REPORT DATE: October 1995

TYPE OF REPORT: Annual

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY *(Leave blank)*	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	October 1995	Annual 30 Sep 94 - 29 Sep 95

4. TITLE AND SUBTITLE
Statistical Methods for Analyzing Time-Dependent Events in Breast Cancer Chemoprevention Studies

5. FUNDING NUMBERS
DAMD17-94-J-4332

6. AUTHOR(S)
Dr. George Y. C. Wong

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Strang Cancer Prevention Center
New York, New York 10021

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

The overall aim of our research proposal is the statistical investigation of a non-parametric estimate which we call redistribution-to-the-inside estimator (RTIE) for the survival function $S(t) = Pr(X > t)$, where the survival time X is subject to interval censoring. Our research efforts in the first year have been focused on two aspects of RTIE for interval-censored data satisfying the condition that for any pair of censoring intervals, either they are disjoint, or one is a subset of the other (DI condition). First, we have completed the implementation of a computer program coded in the C language to carry out the RTIE estimation procedure, including a Kaplan-Meier type plotting program for RTIE written in the S+ language. Second, we derive the uniform almost sure limit (strong consistency) of RTIE for any arbitrary S and any arbitrary censoring distribution function G under the DI model.

14. SUBJECT TERMS
DI Interval Censorship, Redistribution-to-the-inside, Consistency , Breast Cancer

15. NUMBER OF PAGES
47

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	Unlimited

NSN 7540-01-280-5500
Standard Form 298 (Rev. 2-89)
Prescribed by ANSI Std. Z39-18
298-102

GENERAL INSTRUCTIONS FOR COMPLETING SF 298

The Report Documentation Page (RDP) is used in announcing and cataloging reports. It is important that this information be consistent with the rest of the report, particularly the cover and title page. Instructions for filling in each block of the form follow. It is important to ***stay within the lines*** to meet ***optical scanning requirements***.

Block 1. Agency Use Only *(Leave blank)*.

Block 2. Report Date. Full publication date including day, month, and year, if available (e.g. 1 Jan 88). Must cite at least the year.

Block 3. Type of Report and Dates Covered. State whether report is interim, final, etc. If applicable, enter inclusive report dates (e.g. 10 Jun 87 - 30 Jun 88).

Block 4. Title and Subtitle. A title is taken from the part of the report that provides the most meaningful and complete information. When a report is prepared in more than one volume, repeat the primary title, add volume number, and include subtitle for the specific volume. On classified documents enter the title classification in parentheses.

Block 5. Funding Numbers. To include contract and grant numbers; may include program element number(s), project number(s), task number(s), and work unit number(s). Use the following labels:

C	- Contract	**PR**	- Project
G	- Grant	**TA**	- Task
PE	- Program Element	**WU**	- Work Unit Accession No.

Block 6. Author(s). Name(s) of person(s) responsible for writing the report, performing the research, or credited with the content of the report. If editor or compiler, this should follow the name(s).

Block 7. Performing Organization Name(s) and Address(es). Self-explanatory.

Block 8. Performing Organization Report Number. Enter the unique alphanumeric report number(s) assigned by the organization performing the report.

Block 9. Sponsoring/Monitoring Agency Name(s) and Address(es). Self-explanatory.

Block 10. Sponsoring/Monitoring Agency Report Number. *(If known)*

Block 11. Supplementary Notes. Enter information not included elsewhere such as: Prepared in cooperation with...; Trans. of...; To be published in.... When a report is revised, include a statement whether the new report supersedes or supplements the older report.

Block 12a. Distribution/Availability Statement. Denotes public availability or limitations. Cite any availability to the public. Enter additional limitations or special markings in all capitals (e.g. NOFORN, REL, ITAR).

DOD - See DoDD 5230.24, "Distribution Statements on Technical Documents."
DOE - See authorities.
NASA - See Handbook NHB 2200.2.
NTIS - Leave blank.

Block 12b. Distribution Code.

DOD - Leave blank.
DOE - Enter DOE distribution categories from the Standard Distribution for Unclassified Scientific and Technical Reports.
NASA - Leave blank.
NTIS - Leave blank.

Block 13. Abstract. Include a brief *(Maximum 200 words)* factual summary of the most significant information contained in the report.

Block 14. Subject Terms. Keywords or phrases identifying major subjects in the report.

Block 15. Number of Pages. Enter the total number of pages.

Block 16. Price Code. Enter appropriate price code *(NTIS only)*.

Blocks 17. - 19. Security Classifications. Self-explanatory. Enter U.S. Security Classification in accordance with U.S. Security Regulations (i.e., UNCLASSIFIED). If form contains classified information, stamp classification on the top and bottom of the page.

Block 20. Limitation of Abstract. This block must be completed to assign a limitation to the abstract. Enter either UL (unlimited) or SAR (same as report). An entry in this block is necessary if the abstract is to be limited. If blank, the abstract is assumed to be unlimited.

*U.S.GPO:1993-0-358-779

FOREWORD

Opinions, interpretations, conclusions and recommendations are those of the author and are not necessarily endorsed by the US Army.

____ Where copyrighted material is quoted, permission has been obtained to use such material.

____ Where material from documents designated for limited distribution is quoted, permission has been obtained to use the material.

____ Citations of commercial organizations and trade names in this report do not constitute an official Department of Army endorsement or approval of the products or services of these organizations.

____ In conducting research using animals, the investigator(s) adhered to the "Guide for the Care and Use of Laboratory Animals," prepared by the Committee on Care and Use of Laboratory Animals of the Institute of Laboratory Resources, National Research Council (NIH Publication No. 86-23, Revised 1985).

____ For the protection of human subjects, the investigator(s) adhered to policies of applicable Federal Law 45 CFR 46.

____ In conducting research utilizing recombinant DNA technology, the investigator(s) adhered to current guidelines promulgated by the National Institutes of Health.

____ In the conduct of research utilizing recombinant DNA, the investigator(s) adhered to the NIH Guidelines for Research Involving Recombinant DNA Molecules.

____ In the conduct of research involving hazardous organisms, the investigator(s) adhered to the CDC-NIH Guide for Biosafety in Microbiological and Biomedical Laboratories.

Oct 25, 1995

PI - Signature Date

TABLE OF CONTENTS

INTRODUCTION

In clinical follow-up studies, subjects are monitored at regular time intervals for a medical condition. It is often the case that an event under observation can take place in between two successive visits, and it may not be possible for the subject to know the time to such event exactly. For example, consider the situation in which a group of women at high risk for breast cancer is asked to take a chemopreventive substance for a fixed time period. At the end of the period, each participating woman is required to submit a blood or urine sample at regular intervals in order to monitor the level of a validated intermediate biomarker. Let X denote the time from cessation of use of the agent to the loss of its protective effect qualified as a return to baseline value of the biomarker. If a woman submits a sample for assay on a daily basis, the value of X can be observed exactly, unless the protective effect is still present by the time the study is terminated so that X is right-censored in the usual sense of survival analysis. In practice, however, the follow-up interval can be a week or longer; therefore the exact value of X is generally unknown but is known to lie between the time points L and R, where L is the number of days from cessation of agent intake to the last time the sample was assayed and the protective effect was still present, and R is the number of days from cessation of agent intake to the most recent time the sample was assayed. If the protective effect is still present, then R takes the value infinity. In any case, when the value of X is only known to lie between (L, R), we say that X is censored in the interval (L, R). Therefore the observed data consist of either censoring intervals (L, R) or exact observations $X = L = R$.

We consider nonparametric estimation of the distribution function $F(t)$ of a real-valued random variable X (or its survival function $S(t) = 1 - F(t)$, where $F(t) = \mathrm{P}\{X \leq t\}$), when the sample data are incomplete due to restricted observation brought about by interval censoring.

At present, there are only two estimation procedures of S for interval-censored data that are generalized maximum likelihood estimates (GMLE) in the sense of Kiefer and Wolfowitz [1]. The first one is due to Peto [2] and makes use of the Newton-Ralphon algorithm. The second is due to Turnbull [3] and makes use of a self-consistent algorithm. In both cases, there is no closed form expression for the estimator and the algorithm is sample size limiting.

In the first year of our research, we have focused our attention on interval-censored data that satisfy a condition which we call DI condition: data $\{L_1, R_1\}, ..., \{L_n, R_n\}$ are said to satisfy DI condition if given any two censoring intervals, (L_i, R_i) and (L_j, R_j), either they are disjoint or one is a subset of the other. In a clinical study in which every subject has the same follow-up schedule, say at time point a_1, a_2, ..., a_k, then $\{L, R\} = \{0, a_1\}$, or $\{a_i, a_{i+1}\}$ or $\{a_i, \infty\}$, and hence such interval-censoring data will satisfy Condition DI.

Under the DI interval-censorship model, we extend Efron's [4] redistribution-to-the-right idea for right-censored data and propose a redistribution-to-the-inside (RTI) method to yield a nonparametric estimator of $S(t)$ which we call redistribution-to-the-inside estimator (RTIE), denoted by $\hat{S}_I$. Such an estimate has a closed form expression and can be quickly calculated for interval-censored data of any dimension.

In our first year, we have accomplished two important tasks for $\hat{S}_I$:

1. We have implemented a computer program coded in the C language to carry out the RTI procedure, including a Kaplan-Meier [5] type plotting program written in the S+ language for displaying $\hat{S}_I(t)$.
2. We have proved the important result that $\hat{S}_I$ is strongly consistent.

Two completed manuscripts, one pertaining to task (1) and general properties of $\hat{S}_I$ for DI data, and the other pertaining to task (2), are being prepared for submission to peer-reviewed statistical journals. They are included in the Appendix as part of our first year report.

BODY

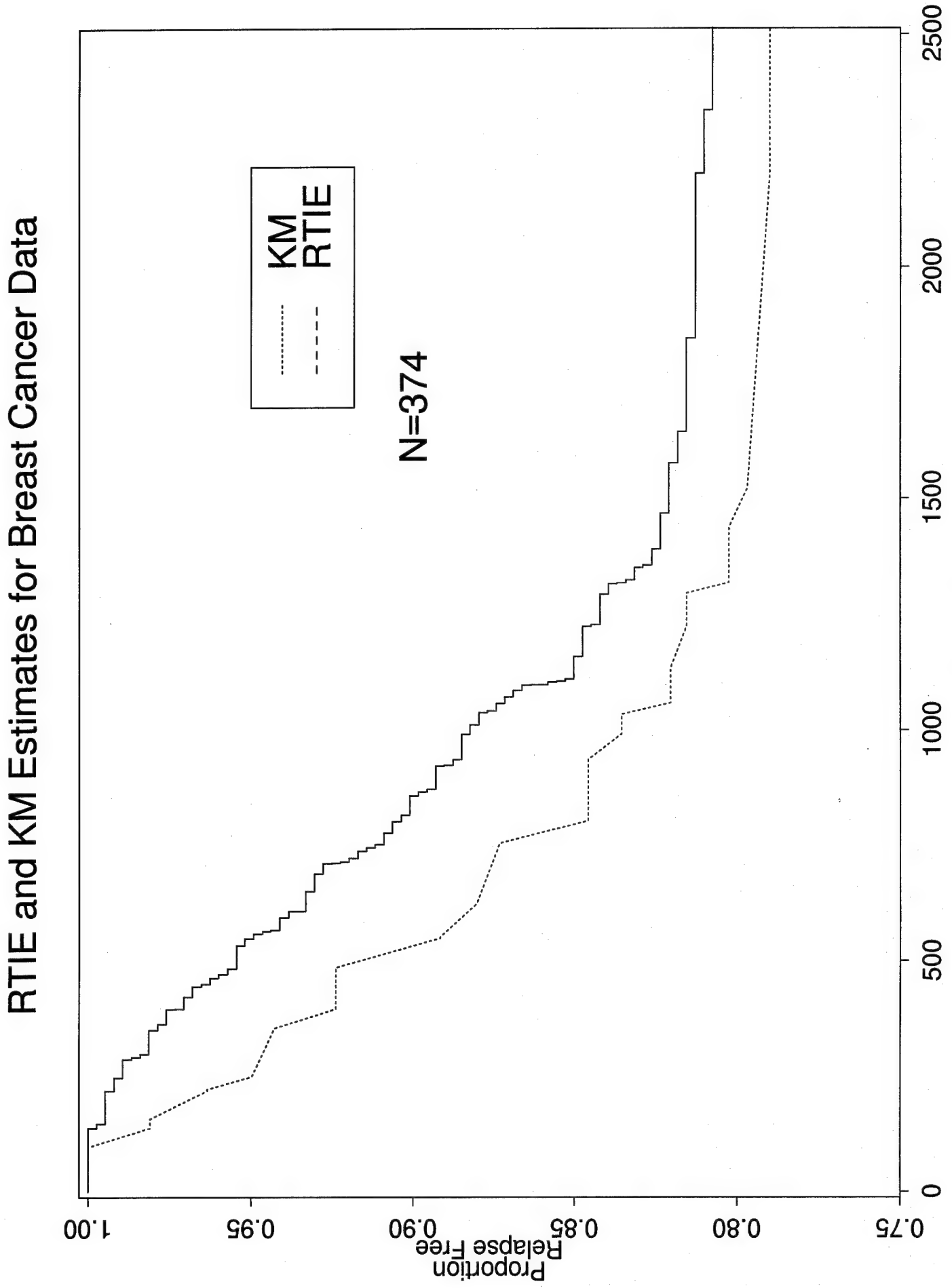
RTI METHOD We here present the idea of our RTI method and the computer program to calculate the RTIE $\hat{S}_I$. Denote $\delta_i = 1[L_i = R_i]$, where $1[A]$ is the indicator function of a set A. For convenience, we first assume that there are no ties in the L_i's and in the R_i's. Let $L_{(i)}$ be the i-th smallest order statistic among the L's and let $\delta_{(i)}$ be the δ_j associated with $L_{(i)}$, so are $X_{(i)}$ and $R_{(i)}$ $i = 1, ..., n$. An observation is said to be a *complete observation* (CO) in an interval, (l, r), if either it is an exact observation which is included in (l, r); or it is a censoring interval which is contained in (l, r).

Although the evaluation of $\hat{S}_I$ does not require any intensive and expensive numerical computing, it does become tedious when the sample size is large. We have implemented our first version of a computer algorithm to calculate $\hat{S}_I$ written in language C. The main portion of the program is given in the following.

We also include a Kaplan-Meier type plot for $\hat{S}_I$ from relapse free survival data from $n = 374$ women with primary stages I, II breast cancer treated by surgery. The corresponding usual Kaplan-Meier estimate treating the interval-censored data as right-censored data is also plotted for comparison purpose.

```
      MAIN PROGRAM FOR RTI PROCEDURE
/* This is the routine to compute estimates of a distribution */
/* interval-censored data  are input from data.in file*/
/* Three estimates are computed here depending on the selection */
/* in the para.in files */
/* There are two more files: rtie.h and util.c    */
/* output file name is given in para.in    */

#include        <stdio.h>
#include        <stdlib.h>
#include        <math.h>
#include        <malloc.h>
#include        "rtie.h"
#include        <time.h>
int     END;
float   INV_POW;
float   para_gen[9][2];
float   **xy;

main()  {
   int   type[3],  simu_switch,  no_qp, no_data;
   int   i, j, n, power();
   int   end1,  Ipara[4], *index;
   int   nout, k, NR, DF;
 float   R[100];
 float   **a, **b, para_in[4][2];
 float   c, xlim, eta;
 float   x[20], nF[20], pF[20], sF[20];
 void    l_qsort();
 void    r_qsort();
 void    swap();
 void    r_swap();
 void    s_trans();
 void    data_form();
 void    weight();
 void    print_info(), read_data(), print_curve(), print_percentile();
 void    print_cdf();
 float   **dmatrix();
 char    ofname[80], multi_fp[80], junk[80];
time_t  tvec;
 FILE    *inf1, *inf2;
 float   *F;

   for (i=0; i< 9; i++)
   for (j=0; j< 2; j++)
   para_gen[i][j] = 1.;

   /* open 1 input data files            */
   inf1 = fopen("para.in", "r");

   /* Read the parameter input file */
   for (i=0; i< 12; i++)
   fscanf(inf1, "%s%*[^\n]", ofname);
   fscanf(inf1, "%d%*[^\n]", &simu_switch);
   fscanf(inf1, "%d%*[^\n]", &END);
   for (i=0; i< 4;i++) {
   fscanf(inf1, "%d", &type[i]);
   fscanf(inf1, "%lf %lf%*[^\n]",&para_in[i][0],&para_in[i][1]);
   }
   fscanf(inf1, "%d%*[^\n]", &Ipara[0]);
   fscanf(inf1, "%d%*[^\n]", &Ipara[1]);
   fscanf(inf1, "%d%*[^\n]", &Ipara[2]);
   fscanf(inf1, "%d%*[^\n]", &Ipara[3]);
   fscanf(inf1, "%s%*[^\n]", ofname);
   fclose(inf1);
   a = (float**) dmatrix(0,1,0,END);
```

```
    F = (float * ) malloc((unsigned) (END * sizeof (float)));
    index  = (int* ) malloc((unsigned) (END * sizeof (int)));

    sprintf(multi_fp, "%s", ofname);
    output = fopen(multi_fp, "w");
    print_info(simu_switch, END, type, para_in, Ipara);
    for (i = 0; i < END; i++) {
          a[0][i] = 0.0;
          a[1][i] = 0.0;
          F[i] = 0.;
          index[i] = 0;
     }
     time(&tvec);
     fprintf(output, "TIME:%s\n", ctime(&tvec));

       /* open  input data file */
       inf2 = fopen("data.in", "r");
       fscanf(inf2, "%s%*[^\n]", junk);
if (simu_switch == 0) {
       for (i=0; i<END ; i++)
       fscanf(inf2, "%lf %lf%*[^\n]", &a[0][i], &a[1][i]);
}

if (simu_switch > 0) {
       i = 0;
       while (i < END) {
       read_data(a,&i,inf2,Ipara);
       }
}

       fclose(inf2);
       for (i=0; i<END ; i++)
                   if (a[0][i]==a[1][i])
                       a[1][i]=0.0;

       if (simu_switch >= 2)  {
       data_form(a, simu_switch); /*use other two approaches*/
       }

       l_qsort(a, END, 0, END-1);
       i=0;
       while (a[0][i]==NEGATIVE & a[1][i]>0.0) i++;
       end1=i;
       r_qsort(a, end1, 0, end1-1);
       s_trans(a, end1);

       end1 = 0;

       weight(a, index, &end1, F);
       print_cdf(a, F, index, &end1, Ipara);

      j = 0;
      for (i=0; i < 19; i++)
      {    while( (x[i] >= a[0][index[j]])  && (j <= end1))
           j++;
           nF[i] += F[j-1];
           sF[i] += F[j-1] * F[j-1];
      }

  time(&tvec);
  fprintf(output, "TIME:%s\n", ctime(&tvec));
}
```


RTIE and KM Estimates for Breast Cancer Data
KM
RTIE
N=374
Proportion Relapse Free
1.00
0.95
0.90
0.85
0.80
0.75
0
500
1000
1500
2000
2500
Time in Days

CONSISTENCY. Under the DI model, we have proved the important result regarding the consistency of $\hat{S}_I$ as a nonparametric estimator of S for interval-censored data. Let us define

$$\mathcal{O} = \{t; t \notin [\tau_l, \tau_r]\},\ \tau_l = \inf\{t; \mathrm{P}\{L < t < R\} = 1 \text{ or } t = +\infty\},$$

$$\tau_r = \min\{\sup\{t; \mathrm{P}\{L < t < R\} = 1\}, +\infty\}.$$

We prove that for any F and censoring distribution function G,

$$\lim_{n\to\infty} \sup_{t\in\mathcal{O}} |\hat{S}_I(t) - S(t)| = 0 \text{ a.s.}$$

If $\tau_l = +\infty$, then $\mathcal{O} = [0, \infty)$. Otherwise,$\mathcal{O}$ is either $[0, \tau_l)$ (right-censorship models) or (τ_r, ∞) (left-censorship models) or $[0, \tau_l) \cup (\tau_r, \infty)$, where $0 < \tau_l < \tau_r < \infty$. Since there are no observations within the interval (τ_l, τ_r) (w.p.1), thus $S(t)$ is not estimatable for $t \in (\tau_l, \tau_r)$.

CONCLUSIONS

As we point out in INTRODUCTION, interval-censored data are commonly encounted in breast cancer follow-up studies and there has been a lack of a computationally feasible statistical procedure for estimating the survival function S even for studies with moderate sample sizes. In our first year of research, we have completed a computer program that can quickly evaluate the nonparametric estimator $\hat{S}_I$ which we propose, and produce a Kaplan-Meier type plot as part of the program. In the BODY section, our program quickly produces the $\hat{S}_I$ plot for overall relapse free survival for interval-censored data from 374 women with stages I, II breast cancer after treatment by surgery. As can be seen from the plot, there is

an appreciable difference between the usual Kaplan-Meier estimator and our $\hat{S}_I$ estimator. The strong consistency that we have established for $\hat{S}_I$ under the DI model is a significant statistical result. We now can reassure users of $\hat{S}_I$ that the estimated value of S will be close to the true value when sample size is moderate.

Our immediate research goals for the second year are to extend the results established here to the case of non-DI data. Specifically, we will extend the RTI method to obtain the counterpart of $\hat{S}_I$ for non-DI data. Then we will investigate conditions under which the corresponding $\hat{S}_I$ can be GMLE and can be consistent. We expect these non-DI extensions to be statistically fairly chanllenging. However, they are obviously very important results, because the majority of interval-censored data in real applications are likely to be non-DI in nature.

REFERENCES

[1] Kiefer, J and Wolfowitz, J. (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *Ann. Math. Statist.* 27, 887-906.

[2] Peto, R. (1973). Experimental survival curves for interval-censored data. *Appl. Statist.* 22, 86-91.

[3] Turnbull, B. W. (1976). The empirical distribution function with arbitrary grouped, censored and truncated data. *J. Roy. Statist. Soc. Ser.* B, 38, 290-295.

[4] Efron, B (1967). The two sample problem with censored data. *Fifth Berkeley Symposium on Mathematical Statistics.* University of California Press, 831-853.

[5] Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *J. Amer. Stat. Assoc.*, 53, 457-481.

Estimation of a Survival Function With Interval-Censored Data under the DI Model

Qiqing Yu * and George Y. C. Wong **

Dept. of Appl. Math. and Statist.,
State University of New York at Stony Brook, NY 11794, USA
and
Strang Cancer Prevention Center,
Cornell University Medical College, 428 E 72nd Street, NY 10021, USA

AMS 1991 subject classification: Primary 62 G05; Secondary 62 A10.

Key words and phrases: interval censorship, redistribution-to-the-inside estimator, double censorship, generalized maximum likelihood estimator, nonparametric estimation.

Summary: We consider nonparametric estimation of a survival function with interval-censored data which satisfy the condition that for any pair of censoring intervals, either they are disjoint or one is a subset of the other. Extending Efron's (1967) idea of redistribution-to-the-right method for deriving the Product-limit estimator (PLE), we propose a redistribution -to-the-inside method which yields an estimate of the survival function, given by a simple, explicit expression. The expression reduces to the PLE under the right censorship model or the left censorship model. The new estimator is shown to be the generalized maximum likelihood estimator of the survival function in the sense of Kiefer and Wolfowitz (1956), and hence is self-consistent in the sense of Turnbull (1976). Extension of the RTI method to the general interval censorship model is also discussed.

1. Introduction. In clinical follow-up studies, subjects are monitored at regular time intervals for a medical condition. It is often the case that an event under observation can take place in between two successive visits, and it may not be possible for the subject to know the time to such event exactly. For example, consider the situation in which a group of women at high risk for breast cancer is asked to take a chemopreventive substance for a fixed time period. At the end of the period, each participating woman is required to submit a blood or urine sample at regular intervals in order to monitor the level of a validated intermediate biomarker. Let X denote the time from cessation of use of the agent to the loss of its protective effect qualified as a return to baseline value of the biomarker. If a woman submits a sample for assay on a daily basis, the value of X can be observed exactly, unless the protective effect is still present by the time the study is terminated so that X is right-censored in the usual sense of survival analysis. In practice, however, the follow-up interval can be a week or longer; therefore the exact value of X is generally unknown but is known to lie between the time points L and R, where L is the number of days from

* Partially supported by NSF Grant DMS-9402561 and DAMD17-94-J-4332.
** Partially supported by DAMD17-94-J-4332.

cessation of agent intake to the last time the sample was assayed and the protective effect was still present, and R is the number of days from cessation of agent intake to the most recent time the sample was assayed. If the protective effect is still present, then R takes the value infinity. In any case, when the value of X is only known to lie between (L, R), we say that X is censored in the interval (L, R). Therefore the observed data consist of either censoring intervals (L, R) or exact observations $X = L = R$.

We consider nonparametric estimation of the distribution function $F(t)$ of a real-valued random variable X (or its survival function $S(t) = 1 - F(t)$, where $F(t) = \mathrm{P}\{X \leq t\}$), when the sample data are incomplete due to restricted observation brought about by interval censoring.

Interval-censored observations consist of vectors $\{L_1, R_1\}$, ..., $\{L_n, R_n\}$, where $L_i \leq R_i$, $i = 1, ..., n$. We assume that these observations are i.i.d. from the population $\{L, R\}$. An observation is said to be *exact* if $L = R = X$, and is called a *censoring interval* if $R - L > 0$. A censoring interval is said to be *empty* if it does not contain exact observations or other censoring intervals (L_i, R_i). Interval-censored data are said to be from a DI interval-censorship model if observations $\{\{L_k, R_k\};\ k = 1, ..., n\}$ satisfy

Condition DI (Disjoint or Included): Given any two censoring intervals, (L_i, R_i) and (L_j, R_j), either they are disjoint or one is a subset of the other.

To illustrate, consider the following two data sets.

$$\big(_1\ \big(_2\ \big)_1\ \big)_2, \tag{1.1}$$

where $(_i$ stands for L_i and $)_i$ stands for R_i, $i = 1, 2$, that is, $L_1 < L_2 < R_1 < R_2$;

$$\big(_1\ \big)_1 \left(\quad \Big(_3\ \big(_4\ \ \big)_4\ \Big)_3 \quad \right)_2 . \tag{1.2}$$

Data set (1.1) does not satisfy Condition DI, whereas Data set (1.2) does. Note that the familiar right-censored data satisfy Condition DI, with $R = +\infty$ if $L < R$, since $(x, +\infty) \supset (y, +\infty)$ if $x < y$. Similarly, the left-censored data also satisfy Condition DI with $L = 0$ if $L < R$ and with half-closed and half-open censoring intervals $[0, R)$.

In a clinical study in which every subject has the same follow-up schedule, say at time point a_1, a_2, ..., a_k, then $\{L, R\} = \{0, a_1\}$, or $\{a_i, a_{i+1}\}$ or $\{a_i, \infty\}$, and hence such interval-censoring data will satisfy Condition DI.

There is only one set of nested censoring intervals in Data set (1.2). Since the right-censored observations form a unique set of nested censoring intervals, it happens that treating empty censoring intervals as exact observations, Data set (1.2) is topologically equivalent to a set of right-censored data: $X_1 < X_2^+ < X_3^+ < X_4$, where X_i^+ stands for a right-censored observation. However, not all DI data are topologically equivalent to right-censored data. For example, in the following DI data,

$$\big(_1\ \big)_1 \left(\quad \big(_3\ \big)_3\ \Big(_4\ \Big(_5\ \big(_6 \big(_7\ \big(_8\ \big)_8\ \big)_7\ \big)_6 \quad \big(_9\ \big(_{10}\ \big)_{10}\ \big)_9\ \Big)_5\ \Big)_4\ \big(_{11}\ \big)_{11} \quad \right)_2\ \big(_{12}\ \big)_{12}, \tag{1.3}$$

there are two sets of nested censoring intervals:

$$\Big(_2\ \big(_4\ \big(_5\ (_6(_7\ (_8\ )_8\ )_7\ )_6\ \ )_5\ \big)_4\ \Big)_2 \text{ and } \Big(_2\ \big(_4\ \big(_5\ \ (_9\ (_{10}\ )_{10}\ )_9\ \big)_5\ \big)_4\ \Big)_2.$$

They are not disjoint (there are common censoring intervals in these two sets), but they cannot form a unique set of nested censoring intervals after excluding empty censoring intervals from the two sets. Thus, Data set (1.3) is not topologically equivalent to right-censored data.

Peto (1973) and Turnbull (1976) consider the problem of obtaining the generalized maximum likelihood estimate (GMLE) of the underlying survival distribution based on interval-censored data (in the sense of Kiefer and Wolfowitz (1956)) using a Newton-Raphson type algorithm and a self-consistent algorithm, respectively. Bacchetti (1990) addresses some extensions of Turnbull's approach. Chang and Yang (1987) and Groeneboom and Wellner (1992) deal with the problem of estimating the underlying survival distribution with doubly-censored data and study the corresponding consistency properties. All these authors, however, do not derive a closed-form expression for their estimator.

It is worth noting that, while the GMLE is unique (Peto (1973)) and is self-consistent (see Tsai and Crowley (1985)), an estimate derived from the self-consistent algorithm may not be unique and thus may not be the GMLE. Gu and Zhang ((1993) page 612) give a counter-example as follows:

Example 1.1. There are two different self-consistent estimates, for a doubly-censored data set with four observations: $V_i = i$, $i = 1, ..., 4$, where V_1 is exact, V_2 is right-censored, and V_3 and V_4 are left-censored. One self-consistent estimate puts mass $2/3$ at 1 and mass $1/3$ at 4, and the other self-consistent estimate puts mass $1/2$ at both 1 and 3. The second estimate is the GMLE, but the first is not.

Because multiple solutions are possible in a self-consistent algorithm, Gu and Zhang (1993) have to add an additional assumption in their theorem for establishing asymptotic normality (Theorem 2), so that the solution $\hat{S}(t)$ from a self-consistent algorithm is indeed the GMLE. Thus it is desirable to find an explicit expression of the GMLE of S. Furthermore, an exact expression of the GMLE $\hat{S}$ will facilitate the establishment of its asymptotic normality and the derivation of its asymptotic variance.

In this paper, we will mainly focus on finding a method to derive an explicit expression for the GMLE under the DI model. Furthermore, we will study the possible extension of the method to the non-DI interval-censored data.

Kaplan and Meier (1958) derive the Product-limit estimator (PLE) for right-censored data. The PLE has a simple expression, in contrast to the numerical solution to the estimate. Efron (1967) shows that the PLE can be obtained through a redistribution-to-the-right (RTR) technique. Efron's idea can be extended to left-censored data, which results in the PLE with a simple expression.

In this paper, we extend Efron's idea and propose a redistribution-to-the-inside (RTI) method to yield an estimate of $S(t)$ with data from a DI model. The new estimate, called the Redistribution-to-the-inside estimate (RTIE), has an explicit expression (see (4.3)). We show in this paper that under the DI model the RTIE is indeed the GMLE. Thus, in this

case, it is the closed form solution to the limit of Newton-Raphson algorithm studied by Peto (1973) and a solution to the limits of the self-consistent algorithm proposed by Turnbull (1976). As a consequence, it is self-consistent according to the definition given in Turnbull (1976). In particular, it reduces to the PLE under the right censorship model and under the left censorship model. Thus, the RTIE unifies the expressions of the PLE with right-censored data and left-censored data.

The motivation for studying the DI model is to find the explicit expression of the GMLE for general interval-censored data, in particular, for a non-DI data set. The RTI method for the DI model may provide some insight on attacking this problem. In this paper, we modify the RTI method for non-DI data. Such an estimator also has an explicit expression. We further show that for a special class of non-DI data the estimate derived from such a modified RTI method is the GMLE. It is worth mention that the data set in Example (1.1) is a non-DI data set and applying our modified RTI method results in the GMLE too.

The RTI method can be implemented as an n-step algorithm, where n is the number of observations. However, it is not a special case of the self-consistent algorithm. The RTI method uniquely defines an estimate; the self-consistent algorithm may result in different estimates depending on the starting points. The RTI method takes no more than n steps; the self-consistent algorithm is an iterative algorithm which stops whenever the error is within a tolerance.

In section 2, we propose the RTI method. In section 3, we show that the new estimator is a GMLE under the DI Model. In section 4, we give a simplified explicit expression of the new estimator under the DI Model.

2. RTI Method. In this section, we will propose a method, which extends Efron's (1967) RTR technique for obtaining the PLE under the right censorship model. We assume that the observations satisfy Condition DI. Denote $\delta_i = 1[L_i = R_i]$, where $1[A]$ is the indicator function of a set A. For convenience, we first assume that there are no ties in the L_i's and in the R_i's. Let $L_{(i)}$ be the i-th smallest order statistic among the L's and let $\delta_{(i)}$ be the δ_j associated with $L_{(i)}$, so are $X_{(i)}$ and $R_{(i)}$ $i = 1, ..., n$. An observation is said to be a *complete observation* (CO) in an interval, (l, r), if either it is an exact observation which is included in (l, r); or it is a censoring interval which is contained in (l, r).

Before we give an estimator of $S(t)$, it is interesting to look at the PLE $\hat{S}_{PL}(t)$ under the right censorship model and the left censorship model.

Under the right censorship model, the observations are (L_1, δ_1),..., (L_n, δ_n) and $R_i = +\infty$ if $L_i < R_i$. The PLE is

$$\hat{S}_{PL}(t) = \prod_{L_{(i)} \leq t} (1 - \frac{\delta_{(i)}}{n - i + 1}). \tag{2.1}$$

Efron's (1967) introduced the RTR method to obtain the PLE: First put mass $1/n$ to each observation L_k. Consider the smallest censoring time $L_{(i)}$. Since a death did not occur at $L_{(i)}$, but somewhere to the right of it, it is reasonable to redistribute $1/n$, the mass at $L_{(i)}$, equally among all observations to the right of $L_{(i)}$ (it can be viewed as to the inside of $(L_{(i)}, R_{(i)})$). Now consider the next censored time, say $L_{(j)}$ $(j > i)$; redistribute $\frac{1}{n} + \frac{1}{n(n-i)}$

among all observations to the right of $L_{(j)}$ (it can be viewed as to the inside of $(L_{(j)}, R_{(j)})$). Treating the other censored times similarly results in the PLE as in (2.1).

On the other hand, the product limit estimator of the distribution function $F(t)$ for the left-censored data can also be obtained by the redistribution-to-the-left (RTL) method.

It can be verified that the data from the left censorship model or right censorship model satisfy Condition DI. Then from the interval-censored data point of view, both redistribution methods can be unified as the redistribution-to-the-inside method. Using this method, we can obtain an estimator $\hat{S}_c(t)$ of $S(t)$ under a more general interval censorship model. We start with the following example to help illustrating the idea of the RTI method.

Example 2.1. Suppose that we have the following 6 observations:

$$\Big(_1 \quad (_2 \quad)_2 \quad (_3 \quad X_4 \quad (_5 \quad)_5 \quad)_3 \quad \Big)_1 \quad (_6 \quad)_6$$

i.e., $L_1 < L_2 < R_2 < L_3 < L_4 = R_4 < L_5 < R_5 < R_3 < R_1 < L_6 < R_6$ ($L_1 \geq 0$ and $R_6 \leq +\infty$). The data satisfy condition DI. Let $p_1, ..., p_6$ be the weights on the observations $\{L_1, R_1\}, \cdots, \{L_6, R_6\}$, respectively. We will derive p_i's in 7 steps:

0. Assign each of the 6 observations weight $1/6$, *i.e.*, $p_i^{(0)} = 1/6$;
1. Since $\{L_1, R_1\}$ is a nonempty censoring interval, *i.e.*, the event occured somewhere inside (L_1, R_1), it is reasonable to redistribute its weight $p_1^{(0)} = 1/6$ to its inside (unless it is an empty interval), that is, to its 4 CO's $\{L_2, R_2\}, \{L_3, R_3\}, \{L_4, R_4\}$ and $\{L_5, R_5\}$ (thus each has $\frac{1}{4} \cdot \frac{1}{6}$ additional weight). Then $p_1^{(1)} = 0$, $p_2^{(1)} = \cdots = p_5^{(1)} = \frac{1}{6}(1+\frac{1}{4}) = \frac{5}{24}$ and $p_6^{(1)} = 1/6$;
2. Since $\{L_2, R_2\}$ is an empty censoring interval, there is no CO inside (L_2, R_2). $p_i^{(2)}$'s remain the same as in the last step;
3. Since $\{L_3, R_3\}$ is a nonempty censoring interval, redistribute its weight $p_3^{(2)} = \frac{1}{6}(1+\frac{1}{4})$ to its 2 CO's $\{L_4, R_4\}$ and $\{L_5, R_5\}$. Thus $p_1^{(3)}$, $p_2^{(3)}$ and $p_6^{(3)}$ remain the same as in the last step and $p_3^{(3)} = 0$, $p_4^{(3)} = p_5^{(3)} = \frac{1}{6}(1+\frac{1}{4})[1+\frac{1}{2}] = \frac{5}{16}$;

k. Since L_4, L_5 and L_6 are either an exact observation or an empty censoring interval, no change is made on $p_i^{(k)}$'s, $k = 4, 5, 6$.

The values $p_i^{(i)}$, $i = 1, ..., 6$, *i.e.*, $(0, \frac{5}{24}, 0, \frac{5}{16}, \frac{5}{16}, \frac{1}{6})$ are the solution to $(p_1, ..., p_6)$; and

$$\hat{S}_c(t) = \begin{cases} 1 & \text{if } t \in [0, R_2) \\ 19/24 & \text{if } t \in [R_2, X_4) \\ 23/48 & \text{if } t \in [X_4, R_5) \\ 1/6 & \text{if } t \in [R_5, R_6) \\ 0 & \text{if } t \geq R_6 \end{cases} \tag{2.2}$$

is the estimate resulting from the RTI method.

From the example, we can see that the method always redistributes the original weight $1/n$ on an empty interval to the *inside* of the same interval, but not to the outside of the interval. For example, the observation $\{L_6, R_6\}$ is to the right of L_1, same as $\{L_2, R_2\}$. However, it is not inside (L_1, R_1), but $\{L_2, R_2\}$ is. The weight on the nonempty interval $\{L_1, R_1\}$ is not redistributed to all the observations to the right of L_1. Thus, it is not a redistribution-to-the-right method or an RTL method.

The formal statement of the method is as follows: We first determine the weight p_i assigned to the i-th ordered observation $(L_{(i)}, R_{(i)})$, then the estimator of $S(t)$ can be constructed easily. Starting from the left of $L_{(1)} \leq \cdots \leq L_{(n)}$, with initial weight $1/n$ for each observation, we derive p_i's in n steps. At the k-th step, if the k-th ordered observation is an exact one or an empty censoring interval, do not make any change to the weights determined at the last step. Otherwise, distribute the weight assigned at the last step to the k-th ordered observation, which is a nonempty censoring interval, to the CO's in the censoring interval.

The following is the formula to determine $\hat{p} = (p_1, ..., p_n)$. Denote

$$N_k = \begin{cases} \#\{\text{all CO's in } (L_{(k)}, R_{(k)})\} & \text{if } \delta_{(k)} = 0 \\ 0 & \text{otherwise,} \end{cases} \tag{2.3}$$

where $\#A$ is the cardinality of the set A. Let the initial value of p_i be

$$p_i^{(0)} = 1/n, \ i = 1, ..., n. \tag{2.4}$$

At Step k, $k = 1, ..., n$,

$$\begin{cases} p_i^{(k)} = p_i^{(k-1)}, \ i = k, ..., n, & \text{if } N_k = 0 \\ \begin{cases} p_k^{(k)} = 0, & \\ p_i^{(k)} = p_i^{(k-1)} + \frac{p_k^{(k-1)}}{N_k} & \text{if } L_{(i)} \text{ is a CO in } (L_{(k)}, R_{(k)}) \\ p_i^{(k)} = p_i^{(k-1)} & \text{if } L_{(i)} \text{ is not a CO in } (L_{(k)}, R_{(k)}) \end{cases} & \text{if } N_k \geq 1. \end{cases} \tag{2.5}$$

$p_i = p_i^{(i)}$ derived from (2.5) is the weight assigned to the i-th ordered observation by the estimator $\hat{S}_c$, $i = 1, ..., n$.

The estimator $\hat{S}_c$ is a probability measure that assigns positive weight to each exact observation and to each empty censoring interval; and assigns no weight to nonempty censoring intervals. If there are no empty censoring intervals in the data, the estimator of $S(t)$ is $\hat{S}_c(t) = \sum_{L_{(i)} > t} p_i$. Otherwise, there exists some k such that $\delta_{(k)} = 0$ and $N_k = 0$. It is well known that in such case the GMLE is not uniquely determined in the interval $(L_{(k)}, R_{(k)})$ (see Peto (1973)). For convenience, we define that the weight p_k is assigned to $R_{(k)}$. Thus in the latter case, the estimator of $S(t)$ is

$$\hat{S}_c(t) = \sum_{L_{(i)} > t, \delta_{(i)} = 1} p_i + \sum_{R_{(i)} > t, \delta_{(i)} = 0} p_i. \tag{2.6}$$

We call $\hat{S}_c$ the redistribution-to-the-inside estimator (RTIE).

Remark 2.1. The PLE is usually undefined for $t > L_{(n)}$ if $\delta_{(n)} = 0$ with right-censored data and for $t < R^*_{(1)}$ if $\delta^*_{(1)} = 0$ with left-censored data, where $R^*_{(1)}$ is the smallest R_j's and $\delta^*_{(1)}$ is the corresponding δ. Expression (2.6) defines $\hat{S}_c(t)$ everywhere for $t \geq 0$.

Remark 2.2. If there is a tie in the L_i's or R_i's, we break the tie as follows:

1. If $\{L_i, R_i\} = \{L_j, R_j\}$, $i < j$, then suppose that L_i occurs before L_j;

2. If $L_i = L_j$ and $L_j < R_j < R_i$, then suppose that L_i occurs before L_j;
3. If an exact observation and the left endpoint of a censoring interval are equal, *i.e.*, $L_i = X_i = L_j < R_j$, then suppose that X_i occurs before L_j.
4. If $L_j < R_j = L_i$, then suppose that R_j occurs before L_i.

Consequently, if, for example, the sample size is two and (L_1, R_1) and (L_2, R_2) are equal censoring intervals, we define the order statistics as $L_{(1)} = L_1$ and $L_{(2)} = L_2$. Furthermore, we regard $\{L_2, R_2\}$ as a CO of (L_1, R_1), but do not regard $\{L_1, R_1\}$ as a CO of (L_2, R_2).

With the above convention, $\hat{S}_c(t)$ as in (2.6) is well defined even when there are ties in the L_i's or in the R_i's.

3. Generalized MLE. We first define the GMLE. Kiefer and Wolfowitz (1956) suggested that for a given nondominated family of probability measure $\mathcal{P}$ one can define a generalized maximum likelihood estimator as follows: For P_1, P_2 in $\mathcal{P}$, let $f(\vec{x}; P_1, P_2) = \frac{dP_1}{dP_2}(\vec{x})$, the Radon-Nikodym derivative of P_1 with respect to $P_1 + P_2$. If $\vec{x}$ represents the observed data vector, $\hat{P}$ is a GMLE if and only if

$$f(\vec{x}; \hat{P}, P) \geq f(\vec{x}; P, \hat{P}) \text{ for all } P \text{ in } \mathcal{P}. \tag{3.1}$$

It is desirable that the RTIE $\hat{S}_c$ is a GMLE. It turns out that this is true under the DI Model.

Hereafter, we denote the lower case letters the values of the corresponding random variables. As discussed in Tsai and Crowley (1985), the definition of the GMLE, $\hat{P}$, of an unknown probability measure P, reduces to $\hat{P}(\vec{x}) \geq P(\vec{x})$, where

$$P(\vec{x}) = \prod_{i=1}^{n} P\{X = l_{(i)}\}^{\delta_{(i)}} P\{X \in (l_{(i)}, r_{(i)})\}^{1-\delta_{(i)}}, \tag{3.2}$$

X is the random variable with the distribution function $F(t)$ and the lower case letters l_i's are the values of corresponding random variables L_i's. In view of Remark 2.2, without loss of generality (WLOG), we can assume that there exist no ties in L_i's and R_j's. Note that the likelihood $P\{X = l_{(i)}\}^{\delta_{(i)}} P\{X \in (l_{(i)}, r_{(i)})\}^{1-\delta_{(i)}}$, for each observation depends only on the values of $F(t)$ at the $L_{(i)}$ and $R_{(i)}$ (see Peto (1973)). Let P assign probability p_i to $l_{(i)}$ if $\delta_{(i)} = 1$ and to the set $[l_{(i)}, r_{(i)}] \setminus \cup_{j>i}\{[l_{(j)}, r_{(j)}]\}$, if $\delta_{(i)} = 0$, $i = 1, ..., n$, where "\" stands for set minus. Given an i, the likelihood

$$P\{X = l_{(i)}\}^{\delta_{(i)}} P\{X \in (l_{(i)}, r_{(i)})\}^{1-\delta_{(i)}} = p_i^{\delta_{(i)}} (\sum_{j=i}^{i+M_i} p_j)^{1-\delta_{(i)}} = \sum_{j=i}^{i+M_i} p_j, \tag{3.3}$$

where

$$M_i = \begin{cases} \#\{j;\ l_{(j)} < r_{(i)}, j > i\} & \text{if } \delta_{(i)} = 0 \\ 0 & \text{otherwise.} \end{cases} \tag{3.4}$$

Then (3.2) is equal to

$$\mathrm{L} = \prod_{i=1}^{n} \sum_{j=i}^{i+M_i} p_j \qquad (\sum_{i=1}^{n} p_i = 1). \tag{3.5}$$

It is well-known that if a censoring interval (l_j, r_j) is empty, the likelihood function Ł would not change as long as the weight on the interval remains the same (see Peto (1973)). Thus the definition of GMLE for t within the empty censoring interval (L_i, R_i) needs not be unique.

Theorem 1. *Suppose that the interval-censored data satisfy Condition DI. Then the RTIE $\hat{S}_c(t)$ (as in (2.6)) of $S(t)$ is a GMLE.*

Proof: In order to prove the theorem, it suffices to show the following statement:

(Sn) the $(p_1, ..., p_n)$ that maximizes Ł as in (3.5) is the same as the $(p_1^{(1)}, ..., p_n^{(n)})$ determined by (2.4) and (2.5).

We prove it by induction on the sample size n.

When the sample size n equals 1, Ł as in (3.5) is maximized by $p_1 = 1$ and (2.5) yields $p_1^{(1)} = p_1^{(0)} = 1$. Thus the theorem is trivially true.

Now assume that the statement (Sn) is true for all sample sizes $n < m$. We will show that the theorem holds also for $n = m$.

When sample size is m, in view of Remark (2.2), WLOG, we can assume that

$$L_1 < \cdots < L_m. \tag{3.6}$$

Since Condition DI is satisfied, one of the following occurs:

(1) $L_1 < L_i \leq R_i < R_1$, $i = 2, ..., m$.

(2) the data set can be partitioned into 2 disjoint subsets, with n_1 observations in the first subset $(1 \leq n_1 < m)$ and n_2 in the second subset $(n_1 + n_2 = m)$.

By disjoint subsets, we mean $\max\{R_i; i \leq n_1\} \leq \min\{L_i; i > n_1\}$.

We first assume that case (2) is true. Let p be the sum of the weights assigned to the elements of the first subset and q the one to the elements of the second subset. Then $p + q = 1$. Note that the likelihood function

$$Ł = [\prod_{i=1}^{n_1}(\sum_{j=i}^{i+M_i} p_j)] \cdot [\prod_{i>n_1}^{m}(\sum_{j=i}^{i+M_i} p_j)] = [\prod_{i=1}^{n_1}(\sum_{j=i}^{i+M_i} \frac{p_j}{p})] \cdot p^{n_1} \cdot [\prod_{i>n_1}^{m}(\sum_{j=i}^{i+M_i} \frac{p_j}{q}] \cdot q^{n_2}.$$

It is easy to see that Ł is maximized by maximizing its three factors :

$$\prod_{i=1}^{n_1}(\sum_{j=i}^{i+M_i} \frac{p_j}{p}), \quad \prod_{i>n_1}^{m}(\sum_{j=i}^{i+M_i} \frac{p_j}{q}) \text{ and } p^{n_1} \cdot q^{n_2}. \tag{3.7}$$

Let $y_i = p_i/p$, then $\sum_{i=1}^{n_1} y_i = 1$ and $i + M_i \leq n_1$ for $i = 1, ..., n_1$. Thus, the first product can be viewed as the likelihood $\prod_{i=1}^{n_1} \sum_{j=i}^{i+M_i} y_j$ of the interval-censored data with n_1 observations: $\{L_1, R_1\}$, ..., $\{L_{n_1}, R_{n_1}\}$. Since $n_1 < m$, by the induction assumption (Sn), the first product is maximized by $y_i^{(i)}$, $i = 1, ..., n_1$, determined by rule (2.5) for the sample size n_1 (and by substituting $p_i = y_i$ in (2.5)).

Let $z_{i-n_1} = p_i/q$, $i = n_1+1, ..., m$. In a similar manner, it can be shown that the second product in (3.7) is maximized by $z_i^{(i)}$, $i = 1, ..., n_2$ determined by rule (2.5) for sample size n_2 with observations $\{L_{n_1+1}, R_{n_1+1}\}$, ..., $\{L_m, R_m\}$ (and by substituting $p_i = z_i$ in (2.5)).

Since the third product in (3.7) is $p^{n_1}(1-p)^{n_2}$, it is maximized by setting $p = n_1/m$ and $1-p = q = n_2/m$.

To complete the proof for case (2) when the sample size is m, it suffices to verify that

(1) For the same data, the weights $(p_1^{(1)}, ..., p_m^{(m)})$ determined by (2.5) with $n = m$, satisfy $p_1^{(1)} + \cdots p_{n_1}^{(n_1)} = n_1/m$ and $p_{n_1+1}^{(n_1+1)} + \cdots + p_m^{(m)} = n_2/m$.

(2) The weights, $p_i^{(i)}$, $i = 1, ..., m$, determined by (2.5) for sample size $n = m$, satisfy that $y_i = p_i^{(i)}/p$, $i = 1, ..., n_1$, is the weight determined by (2.5) for the sample size $n = n_1$ with observations $\{L_1, R_1\}$, ..., $\{L_{n_1}, R_{n_1}\}$, and $z_{i-n_1} = p_i^{(i)}/q$, $i = n_1+1, ..., m$, is the weight determined by (2.5) for the sample size $n = n_2$ with observations $\{L_{n_1+1}, R_{n_1+1}\}$, ..., $\{L_m, R_m\}$.

We first prove statement (1). Note that none of the observations $\{L_{n_1+1}, R_{n_1+1}\}$, ..., $\{L_m, R_m\}$ is a CO of any of the possible censoring intervals (L_i, R_i), $i \le n_1$. Thus the RTI method will not move any of the original weight $1/m$ on $\{L_i, R_i\}$, $i \le n_1$, to $\{L_j, R_j\}$, $j > n_1$. On the other hand, none of the observations $\{L_1, R_1\}$, ..., $\{L_{n_1}, R_{n_1}\}$ is a CO of any of the possible censoring intervals (L_i, R_i), $i > n_1$. Thus the RTI method will not move any of the original weight $1/m$ on $\{L_i, R_i\}$, $i > n_1$, to $\{L_j, R_j\}$, $j \le n_1$. It follows that $p_1^{(1)} + \cdots + p_{n_1}^{(n_1)} = \sum_{i=1}^{n_1} p_i^{(0)} = \sum_{i=1}^{n_1} 1/m = n_1/m$ and $p_{n_1+1}^{(n_1+1)} + \cdots + p_m^{(m)} = n_2/m$. Thus statement (1) holds.

In the following, we prove statement (2). Let $p_i^{(i)}$, $i = 1, ..., m$, be the values of p_i determined by (2.5) (when $n = m$). Then for $i \le n_1$, multiplying $\frac{m}{n_1}$ on both sides of (2.4) and (2.5) yield:

$$\frac{m}{n_1} p_i^{(0)} = \frac{m}{n_1} 1/n = 1/n_1, \ i = 1, ..., n_1; \tag{3.8}$$

and for $k = 1, ..., n_1$,

$$\begin{cases} \frac{m}{n_1} p_i^{(k)} = \frac{m}{n_1} p_i^{(k-1)}, \ i = k, ..., n_1, & \text{if } N_k = 0 \\ \begin{cases} \frac{m}{n_1} p_k^{(k)} = \frac{m}{n_1} 0, & \\ \frac{m}{n_1} p_i^{(k)} = \frac{m}{n_1} p_i^{(k-1)} + \frac{m}{n_1} \frac{p_k^{(k-1)}}{N_k} & \text{if } L_{(i)} \text{ is a CO in } (L_{(k)}, R_{(k)}) \\ \frac{m}{n_1} p_i^{(k)} = \frac{m}{n_1} p_i^{(k-1)} & \text{if } L_{(i)} \text{ is not a CO in } (L_{(k)}, R_{(k)}) \end{cases} & \text{if } N_k \ge 1. \end{cases} \tag{3.9}$$

Let $y_i^{(k)} = \frac{m}{n_1} p_i^{(k)}$, for possible i and k, then (3.8) and (3.9) yield

$$y_i^{(0)} = 1/n_1, \ i = 1, ..., n_1, \tag{3.10}$$

and for $k = 1, ..., n_1$,

$$\begin{cases} y_i^{(k)} = y_i^{(k-1)}, \ i = k, ..., n_1, & \text{if } N_k = 0 \\ \begin{cases} y_k^{(k)} = 0, & \\ y_i^{(k)} = y_i^{(k-1)} + \frac{y_k^{(k-1)}}{N_k} & \text{if } L_{(i)} \text{ is a CO in } (L_{(k)}, R_{(k)}) \\ y_i^{(k)} = y_i^{(k-1)} & \text{if } L_{(i)} \text{ is not a CO in } (L_{(k)}, R_{(k)}) \end{cases} & \text{if } N_k \ge 1. \end{cases} \tag{3.11}$$

Note that (3.10) and (3.11) are identical to (2.4) and (2.5), respectively, provided that the observations are $\{L_1, R_1\}$, ..., $\{L_{n_1}, R_{n_1}\}$ and $n = n_1$. This proves that the weights, $p_i^{(i)}$, $i = 1, ..., m$, assigned by $\hat{S}_c$ for sample size m, satisfy that $y_i^{(i)} = p_i^{(i)}/p$, $i = 1, ..., n_1$, is the weight assigned by $\hat{S}_c$ for the sample size n_1 with observations $\{L_1, R_1\}$, ..., $\{L_{n_1}, R_{n_1}\}$. In the similar manner, we can show that $z_{i-n_1} = p_i^{(i)}/q$, $i = n_1 + 1, ..., m$, is the weight assigned by $\hat{S}_c$ for the sample size n_2 with observations $\{L_{n_1+1}, R_{n_1+1}\}$, ..., $\{L_m, R_m\}$. This completes the proof of statement (2) and the proof for case (2).

To complete the proof for the case $n = m$, we need to show that $\hat{S}_c$ is the GMLE when case (1) is true, *i.e.*, $(L_j, R_j) \subset (L_1, R_1)$ for all $j > 1$. Note that if the latter is true, the likelihood function (3.5) is equal to

$$Ł = (\sum_{i=1}^{m} p_i) \prod_{j=2}^{m} (\sum_{k=j}^{j+M_j} p_k) \qquad (\sum_{i=1}^{m} p_i = 1).$$

Fixing $p_1 + p_2$, Ł increases by setting $p_1 = 0$. That is,

(B1) the solution of the GMLE for p_1 is 0.

When $p_1 = 0$

$$\begin{aligned} Ł &= (\sum_{i=2}^{m} p_i) \prod_{j=2}^{m} (\sum_{k=j}^{j+M_j} p_k) \quad (\sum_{i=2}^{m} p_i = 1) \\ &= \prod_{j=2}^{m} (\sum_{k=j}^{j+M_j} p_k). \end{aligned}$$

The likelihood is the same as the one for the sample size $m - 1$, with observations (L_2, R_2), ..., (L_m, R_m). Thus

(B2) the solution of the GMLE for $(p_2, ..., p_m)$, is the same as the solution, $(p_{*1}, ..., p_{*m-1})$, of the GMLE with $m - 1$ observations $\{L_2, R_2\}, ..., \{L_m, R_m\}$.

We now show that (2.5) yields (B1) and (B2) too. Note that, for $k = 1$, since $N_1 = m-1$, (2.5) yields

$$p_1^{(1)} = 0, \tag{3.12}$$

which is the same as in (B1). Furthermore, for $k = 1$, (2.5) yields

$$p_i^{(1)} = \tfrac{1}{m} + \tfrac{1}{m}/(m-1) = \tfrac{1}{m-1},\ i = 2, ..., m.$$

This can be viewed as (2.4) with $n = m - 1$ and observations $\{L_2, R_2\}, ..., \{L_m, R_m\}$, say,

$$p_{*i}^{(0)} = \frac{1}{m-1}, \quad i = 1, ..., m-1. \tag{3.13}$$

Then

(C) the $(p_2^{(2)}, ..., p_m^{(m)})$ determined by (2.4) and (2.5) for sample size $n = m$ and $k = 2, ..., m$ is the same as $(p_{*1}^{(1)}, ..., p_{*m-1}^{(m-1)})$, determined by (2.4) (or (3.13)) and (2.5) for sample size $n = m - 1$ (and replacing p_i by p_{*i}), with observations $\{L_2, R_2\}, ..., \{L_m, R_m\}$.

By the induction assumption (Sn), and conclusions (B2) and (C), the solution of the GMLE for $(p_2, ..., p_m)$, is the same as the $(p_2^{(2)}, ..., p_m^{(m)})$ determined by (2.4) and (2.5). It indicates that $\hat{S}_c$ is the same as the GMLE for the data of sample size m. This completes the proof for case (1). It also concludes the proof of statement (Sn) for $n = m$ and the proof of the theorem. □

An estimate $\hat{S}$ with interval-censored data under DI model is self-consistent if it satisfies

$$\hat{S}(t) = \frac{1}{n}\sum_{i=1}^{n} 1[L_i > t] + \frac{1}{n}\sum_{L_i \le t < R_i} \frac{\hat{S}(t) - \hat{S}(R_i)}{\hat{S}(L_i) - \hat{S}(R_i)}.$$

It is worth mention that for doubly-censored data, the corresponding equation is different. A GMLE is a self-consistent estimate. It follows:

Corollary. *The estimator $\hat{S}_c$ is self-consistent.*

4. A simple explicit expression of the RTIE under the DI Model. It is expected that the RTIE $\hat{S}_c$ has a simple expression like (2.1) for the PLE. Under the DI Model, the estimator $\hat{S}_c(t)$ can be expressed in the following form: First note that if $(L_{(i)}, R_{(i)})$ and $(L_{(j)}, R_{(j)})$ $(i < j)$ are two censoring intervals that contain t, then Condition DI implies that $\{L_{(j)}, R_{(j)}\}$ is a CO in $(L_{(i)}, R_{(i)})$. Given t, referring N_k as in (2.3), define

$$N_k^t = \begin{cases} \#\{ \text{ all CO's in } (L_{(k)}, t]\} & \text{if } \delta_{(k)} = 0 \text{ and } L_{(k)} < t < R_{(k)} \\ 0 & \text{otherwise} \end{cases} \tag{4.1}$$

and

$$\beta_k(t) = 1[N_k^t > 0]. \tag{4.2}$$

Then

$$1 - \hat{S}_c(t) = \sum_{i=1}^{n} \frac{1[t \ge R_i]}{n} + \sum_{j=1}^{n} \frac{1}{n}\Big[\prod_{k=1}^{j-1}(1 + \frac{1}{N_k})^{\beta_k(t)}\Big] \frac{N_j^t}{(N_j)^{\beta_j(t)}} \tag{4.3}$$

where $\prod_{k=1}^{0} x_k = 1$ for any x_k and $x^0 = 1$ for all $x \ge 0$. Let $t_1 < \cdots < t_m$ be the indices of all (ordered) censoring intervals that contain t (so that $t_1, ..., t_m$ depend on t) and for which $(L_{(t_m)}, R_{(t_m)})$ is not empty. Then (4.3) equals

$$1 - \hat{S}_c(t) = \sum_{i=1}^{n} \frac{1[t \ge R_i]}{n} + \sum_{j=1}^{m} \frac{1}{n}\Big[\prod_{k=1}^{j-1}(1 + \frac{1}{N_{t_k}})\Big] \frac{N_{t_j}^t}{N_{t_j}}. \tag{4.4}$$

Expression (4.4) is another way to express the idea of the redistribution - to - the - inside method. The first term in (4.4), $\sum_i \frac{1[t \ge R_i]}{n}$, is the fraction of the CO's in $(-\infty, +\infty)$ which are in $(-\infty, t]$, *i.e.*, it is the empirical weight carried by the CO's which are $\le t$. Each of the next m summands in (4.4) has two parts: $\frac{N_{t_j}^t}{N_{t_j}}$ is the fraction of the CO's in the censoring interval $(L_{(t_j)}, R_{(t_j)})$ which are in $(L_{(t_j)}, t]$. The quantity, $\frac{1}{n}\big[\prod_{k=1}^{j-1}(1 + \frac{1}{N_{t_k}})\big]$, is the weight accumulated at the j-th nonempty censoring interval up to the $(j-1)$-th step. Thus the

j-th summand is the weight distributed to the total of the CO's in the interval $(L_{(t_j)}, t]$ by the censoring interval $(L_{(t_j)}, R_{(t_j)})$. This indicates that expression (4.4) is the same as expression (2.6) under the DI Model.

It can be shown that expression (4.4) reduces to (2.1) with right-censored data and reduces to the PLE with left-censored data. Thus expression (4.3) or (4.4) is the unified expression which includes the PLE.

Reference.

* Bacchetti, P. (1990). Estimating the incubation period of AIDS by comparing population infection and diagnosis patterns. *JASA*. 85, 1002-1008.
* Chang, M.N. and Yang, G. (1987). Strong consistency of a self-consistent estimator of the survival function with doubly censored data. *Ann. Statist.* 15, 1536-1547.
* Efron, B (1967). The two sample problem with censored data. *Fifth Berkeley Symposium on Mathematical Statistics.* University of California Press, 831-853.
* Groeneboom, P. and Wellner, J. A. (1992). Information bounds and nonparametric maximum likelihood estimation. Birkhäuser Verlag, Basel.
* Gu, M.G. and Zhang, C-H. (1993) Asymptotic properties of self-consistent estimator based on doubly censored data. *Ann. Statist.* 21. 611-624.
* Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *J. Amer. Stat. Assoc.*, 53, 457-481.
* Kiefer, J and Wolfowitz, J. (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *Ann. Math. Statist.* 27, 887-906.
* Peto, R. (1973). Experimental survival curves for interval-censored data. *Appl. Statist.* 22, 86-91.
* Tsai, W. and Crowley, J. (1985). A large sample study of the generalized maximum likelihood estimators from incomplete data via self-consistency. *Ann. Statist.* 13, 1317-1334.
* Turnbull, B. W. (1976). The empirical distribution function with arbitrary grouped, censored and truncated data. *J. Roy. Statist. Soc. Ser.* B, 38, 290-295.

Strong Consistency of the Generalized MLE of a Survival Function under the DI Model

By Qiqing Yu * and George Y. C. Wong **

Dept. of Appl. Math. and Statist.,
State University of New York at Stony Brook, NY 11794, USA
and
Strang Cancer Prevention Center,
Cornell University Medical College, 428 E 72nd Street, NY 10021, USA

First version 9/8/93; latest version 10/27/94
AMS 1991 subject classification: Primary 62 G05 ; Secondary 62 G20.

Key words and phrases: interval censorship, consistency, generalized MLE, nonparametric estimation.

Abstract: Yu and Wong (1993) propose a redistribution-to-the-inside method to derive an explicit expression for the generalized maximum likelihood estimator of an unknown distribution function F with interval-censored data which satisfy that given any pair of censoring intervals either they are disjoint or one is a subset of the other. We call such model a DI model. Both the right censorship model and the left censorship model are special cases of the DI model. Thus, in the latter cases the expression is exactly the Product-limit estimator. In this paper, we derive the uniformly almost sure limit of the estimator on $[0, +\infty)$ for any arbitrary F and any arbitrary censoring distribution function G under the DI Model.

1. Introduction. We consider nonparametric estimation of the distribution function F of a real-valued random variable X (or its survival function $S(t) = 1-F(t) = P\{X > t\}$), when the sample data are incomplete due to restricted observation brought about by interval censoring.

Suppose that $\{X_1, L_1, R_1\}, ..., \{X_n, L_n, R_n\}$ are i.i.d. random vectors from a population $\{X, L, R\}$ and that X and $\{L, R\}$ are independent. We only observe X if $X \notin (L, R)$; otherwise, we only observe $\{L, R\}$. Denote $G(l, r)$ the joint distribution function of $\{L, R\}$ and $\mathcal{V}$ the set of all the possible values (l, r) of random interval (L, R). Denote

$$\{L^*, R^*\} = \begin{cases} \{X, X\} & \text{if } X \notin (L, R) \\ \{L, R\} & \text{otherwise,} \end{cases} \tag{1.1}$$

and denote $\{L_i^*, R_i^*\}$, $i = 1, ..., n$, in an obvious way. Then $\{L_1^*, R_1^*\}$, ..., $\{L_n^*, R_n^*\}$ are interval-censored observations. An observation is said to be *exact* if $L_i^* = R_i^*$ (in which

* Partially supported by NSF Grant DMS-9402561 and DAMD17-94-J-4332.
** Partially supported by DAMD17-94-J-4332.

case, it equals X_i) and is called a *censoring interval* if $R_i^* - L_i^* > 0$. Interval-censored data are said to be from a DI interval-censorship model if observations $\{\{L_i^*, R_i^*\};\ i = 1, ..., n\}$ satisfy

Condition DI (Disjoint or Included): Given any two censoring intervals, (L_i^*, R_i^*) and (L_j^*, R_j^*), either they are disjoint or one is a subset of the other.

The following are examples that data satisfying condition DI may arise.

Example 1.1. Suppose that a_1, a_2,... are scheduled check-up times for patients of some disease ($0 < a_1 < a_2 < \cdots$ and 0 corresponds to the first visit of each patient). Every patient is followed according to this schedule to monitor the disease status. We either know the exact survival time X of the patient or the patient failed to show up since a scheduled check-up, resulting in an L, which is the time the patient last appeared, thus L takes value $a_i \in \{0, a_1, a_2, ...\}$. In the latter case, either the patient is lost to follow-up so that $R = +\infty$, or it is learned at time a_{i+1} (when he missed the appointment) that the patient died before R so that $R \in (a_i, a_{i+1}]$.

Example 1.2. In a cancer follow-up study, patients are monitored for the status of a clinical outcome, such as relapse or disease progression, at scheduled time points. When the inter-follow-up interval is wide, say a few months, and the outcome status requires a careful objective clinical assessment, it may not be possible to know the exact value of the time-to-event variable X (for instance, time from achievement of a complete response to disease progression as determined by rigorous pathological findings) for some patients. The reason for this is that the event can take place sometime between the last and current visits without the patients noticing any changes until they are examined at the current follow-up. For these patients, their X values are known only to lie in an interval and interval-censored data are obtained. In particular, if the schedule is the same for all patients, DI data are obtained.

Note that the familiar right-censored data satisfy Condition DI, with $R = +\infty$ if $L < R$, since $(x, +\infty) \supset (y, +\infty)$ if $x < y$. Similarly, the left-censored data also satisfy Condition DI with $L = -\infty$ if $L < R$.

Peto (1973) and Turnbull (1976) consider the problem of obtaining the generalized maximum likelihood estimate (GMLE) of the underlying survival distribution based on interval-censored data (in the sense of Kiefer and Wolfowitz (1956)) using a Newton-Raphson type algorithm and a self-consistent algorithm, respectively. Chang and Yang (1987) and Groeneboom and Wellner (1992) deal with the problem of estimating the underlying survival distribution with doubly-censored data and study the corresponding consistency properties. Gu and Zhang (1993) establish the strong uniform consistency, asymptotic normality and asymptotic efficiency of the self-consistent estimator under mild conditions on the distribution of censoring variables with doubly-censored data. All these authors, however, do not derive a closed-form expression for their estimator.

Kaplan and Meier (1958) derive the Product-limit estimator (PLE) for right-censored data. Efron (1967) shows that the PLE can be obtained through a redistribution-to-the-right technique. Extending Efron's idea, Yu and Wong (1993) propose a redistribution-to-the-inside (RTI) technique, which unifies the redistribution-to-the-right technique and the redistribution-to-the-left technique (Gomez *et al.* (1992)), and obtain an estimate with DI interval-censored data. The estimator, called the RTIE and denoted by $\hat{S}_I(t)$, has an

explicit expression and is the GMLE under the DI Model (Yu and Wong (1993)).

Gu and Zhang ((1993) page 612) give an example that a self-consistent algorithm may result in multiple solutions for self-consistent estimate and a self-consistent estimate may not be the GMLE. Thus it is edsirable to find the explicit expression of the GMLE. The motivation for studying the DI model is to find the explicit expression of the GMLE for general interval-censored data, in particular, for a non-DI data set. The RTI method for the DI model may provide some insight on attacking this problem. Yu and Wong (1994) modify the RTI method for non-DI data. The method also results in an explicit expression of the RTIE. They further show that for a special class of non-DI data the estimate derived from the modified RTI method is the GMLE.

Under the DI Model $\hat{S}_I(t)$ is the closed form solution to the limit of Newton-Raphson algorithm studied by Peto (1973) and a solution to the limits of the self-consistent algorithm proposed by Turnbull (1976). As a consequence, it is self-consistent according to the definition given in Turnbull (1976). In particular, it reduces to the PLE under the right censorship model and under the left censorship model. We only consider DI models in this paper.

We derive the almost sure limit of $\hat{S}_I(t)$ uniformly on $[0, +\infty)$ for any arbitrary F and G (see Theorem 4.2 and Remark 5.1). In proofs, we use a real analysis approach. In light of the literature, it is conceivable that if we use a stochastic process approach or martingale approach (see, for example, Gill (1983) or Stute and Wang (1993)), the proof may be shorten. However, using such approach, additional assumptions on F or G are needed. For example, Stute and Wang (1993) use a martingale approach to show that with right-censored data the PLE is strongly consistent uniformly on the interval $t < \tau$, for any arbitrary F and G, provided F and G do not have any discontinuous points in common, where $\tau = \inf\{s;\ F(s) = 1 \text{ or } G(s, +\infty) = 1\}$. Using our approach, we do not have any assumption on F and G. Furthermore, our main results imply a stronger result than Gomez *et al.* (1992) result on the consistency of the PLE with left-censored data: they show that the PLE with left-censored data is strongly consistent uniformly on $t \geq t_0$ for any arbitrary t_0, F and G, provided $F(t_0) > 0$ and $G(+\infty, t_0) > 0$. Our main results imply that the PLE with left-censored data is strongly consistent uniformly on $t \geq \tau_r$ for any arbitrary F and G, where $\tau_r = \inf\{r;\ G(+\infty, r) > 0\}$.

In Section 2, we define the notation. In Section 3, we give a consistency proof when G is discrete. In Section 4, we give the main consistency results (Theorem 4.2). In Section 5, we discuss the almost sure limit on the half real line. Some proofs of lemmas and theorem are put in the Appendix. For the convenience of readers, we give details in proofs, which may be condensed in a future revision. In particular, the Appendix could be deleted in a future revision, since the main idea of the consistency proof for an arbitrary G is in the proof of Theorem 3.1 for a discrete G.

2. Notation and the GMLE. Hereafter, we assume that observations are from a DI Model. Denote $\delta_i = 1[X_i \notin (L_i, R_i)]$, where $1[A]$ is the indicator function of a set A. For convenience, we first assume that there is no tie in the L_i^*'s. Let $L_{(i)}^*$ be the i-th smallest order statistic of L_j^*'s and let $R_{(i)}^*$ be the R_j^* associated with $L_{(i)}^*$, and denote $\delta_{(i)}$, $X_{(i)}$, $L_{(i)}$ and $R_{(i)}$ in a similar manner, $i = 1, ..., n$. An observation $\{L_i^*, R_i^*\}$ is right-censored if $R_i^* = +\infty$ and is left-censored if $L_i^* = -\infty$ and $L_i^* < R_i^*$. A censoring interval is said

to be *empty* if it does not contain exact observations or other censoring intervals (L_j^*, R_j^*). $\{L_i^*, R_i^*\}$ is said to be a *complete observation* (CO) in an interval if either it is an exact observation which belongs to the interval or it is a censoring interval which is a subset of the interval.

The RTI method works as follows. Initial weight $1/n$ is assigned to each observation. Then starting from $\{L_{(1)}^*, R_{(1)}^*\}$, if it is a nonempty censoring interval, i.e., the event occured somewhere inside $(L_{(1)}^*, R_{(1)}^*)$, it is reasonable to redistribute its weight equally to the observations inside the interval; otherwise do not make any change. Treating $\{L_{(i)}^*, R_{(i)}^*\}$, $i = 2, 3, ..., n$, similarly results in the RTIE $\hat{S}_I(t)$.

Write $\hat{F}_I(t) = 1 - \hat{S}_I(t)$. Yu and Wong show that

$$\hat{F}_I(t) = \sum_{i=1}^{n} \frac{1[R_i^* \le t]}{n} + \sum_{j=1}^{n} \frac{1}{n} \Big[\prod_{1 \le k < j} (1 + \frac{1}{N_k})^{\beta_k(t)} \Big] \frac{N_j^t}{(N_j)^{\beta_j(t)}}, \tag{2.1}$$

where

$$N_k = \begin{cases} \#\{\text{all CO's in } (L_{(k)}^*, R_{(k)}^*)\} & \text{if } \delta_{(k)} = 0 \\ 0 & \text{otherwise,} \end{cases}$$

$$\#A \quad \text{is the cardinality of a set} \quad A,$$

$$N_k^t = \begin{cases} \#\{\text{all CO's in } (L_{(k)}^*, t]\} & \text{if } \delta_{(k)} = 0 \text{ and } L_{(k)}^* < t < R_{(k)}^* \\ 0 & \text{otherwise ,} \end{cases} \tag{2.2}$$

$\beta_k(t) = 1[N_k^t > 0]$, $x^0 = 1$ for all $x \ge 0$ and $\prod_{1 \le k < 1} (1 + \frac{1}{N_k})^{\beta_k(t)} = 1$. Under the DI Model, Yu and Wong (1993) show that $\hat{S}_I(t)$ is the GMLE of S. It is worth noting that $\hat{F}_I(t)$ is right continuous, nondecreasing in t and bounded by $[0, 1]$. It is well known (see Peto (1973)) that the GMLE is uniquely defined in terms of weights on exact observations or on empty intervals, but is not uniquely defined for t within an empty censoring interval. However, the definition of the GMLE for t in the following three cases affects its property of the uniformly strong consistency: (1) the largest order statistic $L_{(n)}^*$ is right-censored; (2) $\min_i\{R_i^*\}$ is left-censored; or (3) $(L_{(n)}^*, R_{(n)}^*)$ is an empty censoring interval and $R_{(n)}^* = \min_i\{R_i^*\}$. (See, for example, Yu and Li (1994).) In particular, if case (2) (or case (3)) is true and $\hat{S}_I(t)$, $t < \min_i R_i^*$ (or $t \in [L_{(n)}^*, R_{(n)}^*]$), is defined as in (2.1), then $\hat{S}_I(t)$ is not uniformly strongly consistent (**for $t < \tau$ Delete the following**) for $t \in \mathcal{O}$. Thus we use the following modification.

Remark 2.1. If either case (2) or case (3) occurs, we modify (2.1) as follows: in case (2), $\hat{S}_I(t) = \hat{S}_I(\min_i R_i^*)$ for $t < \min_i R_i^*$; in case (3), define $\hat{S}_I(t)$ to be a right continuous step function with a unique jump at the median of $L_{(n)}^*$ and $R_{(n)}^*$ for $t \in [L_{(n)}^*, R_{(n)}^*]$.

Remark 2.2. If there are ties in the L_i^*'s, neither N_k nor expression (2.1) is well defined. In such cases, we break the ties as follows:

1. If $\{L_i^*, R_i^*\} = \{L_j^*, R_j^*\}$, $i < j$, then suppose that L_i^* occurs before L_j^*;
2. If $L_i^* = L_j^*$ and $L_j^* < R_j^* < R_i^*$, then suppose that L_i^* occurs before L_j^*;
3. If an exact observation and the left endpoint of a censoring interval are equal, i.e., $L_i^* = X_i = L_j^* < R_j^*$, then suppose that X_i occurs before L_j^*.

4. If $L_j^* < L_i^* = R_j^*$, then suppose that R_j^* occurs before L_i^*.

Thus, for example, if the sample size is 2 and the two censoring intervals (L_1^*, R_1^*) and (L_2^*, R_2^*) are equal, we define the order statistics as $L_{(1)}^* = L_1^*$ and $L_{(2)}^* = L_2^*$, and we regard $\{L_2^*, R_2^*\}$ as a CO of (L_1^*, R_1^*), but do not regard $\{L_1^*, R_1^*\}$ as a CO of (L_2^*, R_2^*).

Recall that under the right censorship model, with probability 1 (w.p.1) there are no exact observations for $t > \tau$, thus, people only study the consistency of the PLE for $t \leq \tau$. A similar condition, namely, $t \in \mathcal{O}$, occurs in interval-censored data, where

$$\mathcal{O} = \{t; t \notin [\tau_l, \tau_r]\},\ \tau_l = \inf\{t; \mathrm{P}\{L < t < R\} = 1 \text{ or } t = +\infty\},$$
$$\tau_r = \min\{\sup\{t; \mathrm{P}\{L < t < R\} = 1\}, +\infty\}. \tag{2.2}$$

If $\tau_l = +\infty$, then $\mathcal{O} = [0, \infty)$. Otherwise, $\mathcal{O}$ is either $[0, \tau_l)$ (right-censorship models) or (τ_r, ∞) (left-censorship models) or $[0, \tau_l) \cup (\tau_r, \infty)$, where $0 < \tau_l < \tau_r < \infty$. There are no observations within the interval (τ_l, τ_r) (w.p.1), thus $F(t)$ (or $S(t)$) is not estimatable for $t \in (\tau_l, \tau_r)$. Denote $\overline{A}$ the closure of a set A. We will study the consistency of $\hat{F}_I(t)$ (or $\hat{S}_I(t)$) for $t \in \mathcal{O}$ or on its boundary.

3. Consistency of $\hat{S}_I(t)$ when G is discrete. One of our main results is

$$(1)\ \lim_{n\to\infty} \sup_{t\in\mathcal{O}} |\hat{S}_I(t) - S(t)| = 0 \text{ a.s and } (2)\ \lim_{n\to\infty} \sup_{t\in\mathcal{O}} |\hat{F}_I(t) - F(t)| = 0 \text{ a.s.}, \tag{3.1}$$

for any F and G. For a better presentation, we first prove (3.1) when G is discrete. The proof is very typical in terms of the technique used in this paper and is easy to follow. The proof of the main result is similar to this proof with modification to deal with the complexity arising from relaxing the assumption on G.

Theorem 3.1. *Suppose that $G(l, r)$ is a discrete distribution function. Then (3.1) holds.*

Proof: Note that (1) and (2) in (3.1) are equivalent. There are two summations in the expression of $\hat{F}_I(t)$ (see (2.1)). We will derive their almost sure limits and show that their sum is $F(t)$. It is easy to show that the first summation

$$\lim_{n\to\infty} \sum_{i=1}^{n} \frac{1[R_i^* \leq t]}{n} = F(t) - \mathrm{P}\{L < X \leq t < R)\} \text{ a.s. uniformly for } t \geq 0, \tag{3.2}$$

since $\{R^* \leq t\} = \{X \leq t\} \setminus \{L < X \leq t < R\}$ (see (1.1). Denote $Q_n(t)$ the second summation in expression (2.1), i.e.,

$$Q_n(t) = \sum_{j=1}^{n} \frac{1}{n} [\prod_{1\leq k<j} (1 + \frac{1}{N_k})^{\beta_k(t)}] \frac{N_j^t}{(N_j)^{\beta_j(t)}} \quad (= \hat{F}_I(t) - \sum_{i=1}^{n} \frac{1[R_i^* \leq t]}{n}). \tag{3.3}$$

It follows from (3.2) and (3.3) that

$$\lim_{n\to\infty} \sup_{t\in\mathcal{O}} |\hat{F}_I(t) - F(t)| = \lim_{n\to\infty} \sup_{t\in\mathcal{O}} |Q_n(t) - \mathrm{P}\{L < X \leq t < R\}| \tag{3.4}$$

$$= \lim_{n\to\infty} \sup_{t\in\mathcal{O}} |Q_n(t) - \sum_{i=1}^{m} \mathrm{P}\{X \in (l_i, t]\}\mathrm{P}\{(L, R) = (l_i, r_i)\}|, \tag{3.5}$$

where (l_j, r_j), $j = 1, 2, ..., m$, are all the possible distinct values of $\mathcal{V}$ which satisfy:

$$l_1 \leq \cdots \leq l_j < t < r_j \leq \cdots \leq r_1 \text{ (note that } l_j, r_j \text{ and } m \text{ are functions of } t),$$
$$\mathrm{P}\{X \in (l_j, t]\} > 0,\ j = 1, ..., m \ (m \text{ maybe } 0 \text{ or } +\infty). \tag{3.6}$$

Hereafter, we will show that expression (3.5) equals 0 a.s. (i.e., (3.1) holds). Note that if $m = 0$, then $\mathrm{P}\{L < t < R\} = 0$ and $N_k^t = 0$. Thus $Q_n(t) = 0 = \mathrm{P}\{L < X \leq t < R\}$. Without loss of generality (WLOG), we can assume that $m \geq 1$ a.s. for all t. Since G is discrete, ties may occur in censoring intervals. Given t and (l_j, r_j) which satisfy (3.6), let $q_j^0 = \#\{i;\ (L_i^*, R_i^*) = (l_j, r_j)\}$ (the number of ties at (l_j, r_j)) for $j = 1, ..., m$. By an induction argument on m, it can be shown (see Lemma 6.1 in the Appendix) that

$$Q_n(t) = \sum_{j=1}^{m} \Big\{ [\frac{q_j^0}{n}][\prod_{1 \leq k < j} (\frac{N_{k*} + q_k^0}{N_{k*}})^{\beta_{k*}(t)}][\frac{N_{j*}^t}{(N_{j*})^{\beta_{j*}(t)}}] \Big\}, \tag{3.7}$$

where

$$N_{i*}^t = \#\{\text{all CO's in } (l_i, t]\} \quad (= \#\{k;\ X_k \in (l_i, t]\} - \#\{k;\ X_k \in (l_i, t], L_k^* < t < R_k^*)\}),$$
$$\beta_{i*}(t) = 1[N_{i*}^t > 0], \tag{3.8}$$
$$N_{i*} = \#\{\text{all CO's in } (l_i, r_i)\}\ (= \#\{k; X_k \in (l_i, r_i)\} - \#\{k; X_k \in (l_i, r_i), (L_k, R_k) \supset (l_i, r_i)\}).$$

To derive the limit of $Q_n(t)$, we need to derive the limits of the three factors in (3.7). The three limits will be given in (3.9), (3.12) and (3.13) below. Then we derive the limit of $Q_n(t)$ in (3.15). The derivation is as follows.

Since X and $\{L, R\}$ are independent, we have $\mathrm{P}\{(L^*, R^*) = (l, r)\} = \mathrm{P}\{X \in (l, r)\} \cdot \mathrm{P}\{(L, R) = (l, r)\}$, and uniformly for all possible $l_j < r_j$,

$$\lim_{n\to\infty} \frac{q_j^0}{n} = \lim_{n\to\infty} \frac{\#\{i;\ (L_i^*, R_i^*) = (l_j, r_j)\}}{n} = \mathrm{P}\{X \in (l_j, r_j)\}\mathrm{P}\{(L, R) = (l_j, r_j)\} \text{ a.s..} \tag{3.9}$$

To derive the limits of the other factors in (3.7), note that the first equality in (3.8) yields

$$\lim_{n\to\infty} N_{j*}^t/n \geq \mathrm{P}\{X \in (l_j, t]\} - \mathrm{P}\{X \in (l_j, t]\}\mathrm{P}\{L < t < R\} \text{ a.s.} \tag{3.10}$$

for all j and uniformly for all $t \in \mathcal{O}$. Since $l_j < t < r_j$ by notation in (3.6) and $t \in \mathcal{O}$, we have $\mathrm{P}\{L < t < R\} < 1$. It follows from (3.6), (3.10) and the last inequality that

$$\lim_{n\to\infty} N_{j*}^t/n \geq \mathrm{P}\{X \in (l_j, t]\}[1 - \mathrm{P}\{L < t < R\}] > 0 \text{ a.s.}$$

for all j and for all $t \in \mathcal{O}$. Consequently, $\lim_{n\to\infty} \beta_{j*}(t) = 1$ a.s.. WLOG, we can assume that $\beta_{j*}(t) = 1$. Then the limit of the last factor in (3.7) is given by

$$\lim_{n\to\infty} \frac{N_{j*}^t}{N_{j*}} \quad (= \frac{\lim_{n\to\infty} N_{j*}^t/n}{\lim_{n\to\infty} N_{j*}/n})$$

$$= \frac{\mathrm{P}\{X \in (l_j, t]\} - \mathrm{P}\{X \in (l_j, t], L^* < t < R^*\}}{\mathrm{P}\{X \in (l_j, r_j)\} - \mathrm{P}\{X \in (l_j, r_j), (L, R) \supset (l_j, r_j)\}} \quad \text{(see (3.8))}$$

$$= \frac{\mathrm{P}\{X \in (l_j, t]\} - \sum_{i=1}^{j} \mathrm{P}\{X \in (l_j, t], (L, R) = (l_i, r_i)\} - \sum_{i>j}^{m} \mathrm{P}\{X \in (l_i, t], (L, R) = (l_i, r_i)\}}{\mathrm{P}\{X \in (l_j, r_j)\}[1 - \sum_{i=1}^{j} \mathrm{P}\{(L, R) = (l_i, r_i)\}]}$$

$$= \frac{\mathrm{P}\{X \in (l_j, t]\}}{\mathrm{P}\{X \in (l_j, r_j)\}} - \frac{\sum_{i>j}^{m} \mathrm{P}\{X \in (l_i, t]\}\mathrm{P}\{(L, R) = (l_i, r_i)\}}{\mathrm{P}\{X \in (l_j, r_j)\}[1 - \sum_{i=1}^{j} \mathrm{P}\{(L, R) = (l_i, r_i)\}]} \text{ a.s.} \tag{3.12}$$

uniformly for all $t \in \mathcal{O}$, $j \geq 1$.

The limit of the product in the summand of (3.7) is given by

$$\lim_{n\to\infty} \prod_{1 \leq k < j} \frac{N_{k*} + q_k^0}{N_{k*}} = \prod_{1 \leq k < j} \frac{\mathrm{P}\{X \in (l_k, r_k)\}[1 - \sum_{h<k} \mathrm{P}\{(L, R) = (l_h, r_h)\}]}{\mathrm{P}\{X \in (l_k, r_k)\}[1 - \sum_{h \leq k} \mathrm{P}\{(L, R) = (l_h, r_h)\}]}$$

$$= \frac{1}{1 - \sum_{h<j} \mathrm{P}\{(L, R) = (l_h, r_h)\}} \text{ a.s.} \tag{3.13}$$

uniformly for all $t \in \mathcal{O}$, $j \geq 1$. (3.7), (3.9), (3.12) and (3.13) yield

$$\lim_{n\to\infty} Q_n(t) = \sum_{j=1}^{m} \Big\{ \big[\mathrm{P}\{X \in (l_j, r_j)\}\mathrm{P}\{(L, R) = (l_j, r_j)\}\big] \cdot \frac{1}{1 - \sum_{h<j} \mathrm{P}\{(L, R) = (l_h, r_h)\}}$$

$$\cdot \Big[\frac{\mathrm{P}\{X \in (l_j, t]\}}{\mathrm{P}\{X \in (l_j, r_j)\}} - \frac{\sum_{i>j}^{m} \mathrm{P}\{X \in (l_i, t]\}\mathrm{P}\{(L, R) = (l_i, r_i)\}}{\mathrm{P}\{X \in (l_j, r_j)\}[1 - \sum_{i=1}^{j} \mathrm{P}\{(L, R) = (l_i, r_i)\}]} \Big] \Big\}$$

$$= \sum_{j=1}^{m} \Big\{ \frac{\mathrm{P}\{X \in (l_j, t]\}\mathrm{P}\{(L, R) = (l_j, r_j)\}}{1 - \sum_{h<j} \mathrm{P}\{(L, R) = (l_h, r_h)\}} \tag{3.14}$$

$$- \frac{\mathrm{P}\{(L, R) = (l_j, r_j)\} \sum_{i>j}^{m} \mathrm{P}\{X \in (l_i, t]\}\mathrm{P}\{(L, R) = (l_i, r_i)\}}{[1 - \sum_{h<j} \mathrm{P}\{(L, R) = (l_h, r_h)\}][1 - \sum_{i=1}^{j} \mathrm{P}\{(L, R) = (l_i, r_i)\}]} \Big\} \text{ a.s.}$$

uniformly for all $t \in \mathcal{O}$. In view of (3.5), to prove the theorem it suffices to show that the last expression of (3.14) equals $\sum_{i=1}^{m} \mathrm{P}\{X \in (l_i, t]\}\mathrm{P}\{(L, R) = (l_i, r_i)\}$. That is

$$0 = \sum_{j=1}^{m} \Big\{ \frac{\mathrm{P}\{X \in (l_j, t]\}\mathrm{P}\{(L, R) = (l_j, r_j)\}}{1 - \sum_{h<j} \mathrm{P}\{(L, R) = (l_h, r_h)\}}$$

$$- \frac{\mathrm{P}\{(L, R) = (l_j, r_j)\} \sum_{i>j}^{m} \mathrm{P}\{X \in (l_i, t]\}\mathrm{P}\{(L, R) = (l_i, r_i)\}}{[1 - \sum_{h<j} \mathrm{P}\{(L, R) = (l_h, r_h)\}][1 - \sum_{i=1}^{j} \mathrm{P}\{(L, R) = (l_i, r_i)\}]} \Big\}$$

$$- \sum_{i=1}^{m} \mathrm{P}\{X \in (l_i, t]\}\mathrm{P}\{(L, R) = (l_i, r_i)\} \tag{3.15}$$

$$= \sum_{k=1}^{m} \Big\{ \frac{\mathrm{P}\{X \in (l_k, t]\}\mathrm{P}\{(L, R) = (l_k, r_k)\}}{1 - \sum_{h<k} \mathrm{P}\{(L, R) = (l_h, r_h)\}} - \mathrm{P}\{X \in (l_k, t]\}\mathrm{P}\{(L, R) = (l_k, r_k)\} \tag{3.16}$$

$$- \sum_{1 \leq j < k} \Big[\frac{\mathrm{P}\{(L, R) = (l_j, r_j)\}\mathrm{P}\{X \in (l_k, t]\}\mathrm{P}\{(L, R) = (l_k, r_k)\}}{[1 - \sum_{h<j} \mathrm{P}\{(L, R) = (l_h, r_h)\}][1 - \sum_{i=1}^{j} \mathrm{P}\{(L, R) = (l_i, r_i)\}]} \Big] \Big\},$$

which is proved in Lemma 6.2 (in the Appendix) by showing that each summand in expression (3.16) equals 0. This completes the proof of the theorem. □

4. Main results. In this section, we will prove (3.1) assuming F and G are arbitrary. We will also investigate the consistency of $\hat{S}_I(t)$ (or $\hat{F}(t)$) on the boundary of $\mathcal{O}$. We first establish Lemma 4.1, which reduces the uniformly almost sure convergency to the pointwise almost sure convergency.

Lemma 4.1. *Suppose that $\{F_n\}_{n\geq 1}$ is a sequence of monotone functions on an interval $[a,b)$ and $F(t)$ is a bounded monotone and right continuous function on the interval $[a,b)$. If*

$$\lim_{n\to\infty} F_n(t) = F(t) \ \forall\ t \in [a,b) \text{ and } \lim_{n\to\infty} F_n(t-) = F(t-) \ \forall\ t \in (a,b],$$

where $F(t-) = \lim_{s\uparrow t} F(s)$, then $\lim_{n\to\infty} \sup_{t\in[a,b)} |F_n(t) - F(t)| = 0$.

If F is continuous then the lemma is a well known result. The proof of the lemma is very similar to the proof for a continuous F and is put in the Appendix.

It is easy to verify that $\hat{S}_I(t)$ is a monotone function of t, in view of Lemma 4.1, to prove (3.1), it suffices to show that

$$\text{(S1)}\ \lim_{n\to\infty} \hat{S}_I(t) = S(t) \text{ and (S2) } \lim_{n\to\infty} \hat{S}_I(t-) = S(t-), \tag{4.1}$$

for each $t \in \mathcal{O}$, (S1) holds for $t = \tau_r$ and (S2) holds for $t = \tau_l$. To avoid the complexity on the boundary of $\mathcal{O}$, we first treat the case that $t \in \mathcal{O}$ in Theorem 4.1. On the boundary of $\mathcal{O}$, depending on whether F or G is continuous on the boundary, there is some modification on the proofs, thus we give the proof for each case separately (in Lemmas 4.3, 4.4 and 4.5). However the main idea is similar to the proof of Theorem 3.1.

Theorem 4.1. For any arbitrary F and G, and for any $t \in \mathcal{O}$, (4.1) holds.

Proof: We first show equality (S1) by mimicing the proof of Theorem 3.1. Note that in the latter proof, the arguments up to (3.4) hold for any F and G. Hence, (3.4) implies that it suffices to show that

$$\lim_{n\to\infty} Q_n(t) = \mathrm{P}\{L < X \leq t < R\}.$$

or equivalently, to show that for any large positive integer m, we have

$$\overline{\lim_{n\to\infty}} \left\{Q_n(t) - \mathrm{P}\{L < X \leq t < R\}\right\} \leq O(1/\sqrt{m}) \text{ a.s.} \tag{4.2}$$

$$\overline{\lim_{n\to\infty}} \left\{ - Q_n(t) + \mathrm{P}\{L < X \leq t < R\}\right\} \leq O(1/\sqrt{m}) \text{ a.s..} \tag{4.3}$$

WLOG, we can assume that $(-\infty, 0) \in \mathcal{V}$. Denote $\mathcal{B} = \cup_{(l,r)\in\mathcal{V}}(l,r)$ and $\mathcal{B}^c$ its complement set. If $t \in \mathcal{B}^c$, then $\mathrm{P}\{L < t < R\} = 0$ and thus $Q_n(t)$=0 since $N_k^t = 0$ for all k (see (2.2) and (3.3)). It follows that $\mathrm{P}\{L < X \leq t < R\} = 0 = Q_n(t)$, which implies (4.2) and (4.3). Thus we assume $t \in \mathcal{B}$. To show (4.2) and (4.3), we will mimic the

arguments from (3.5) through (3.14) in the proof of Theorem 3.1. However, there may be uncountably many elements in $\mathcal{V}$ such that (3.6) holds. Thus we modify (3.6) as follows:

Given $t \in \mathcal{B}$ and given a large positive integer m, we can find a finite partition, $\{\mathcal{C}_m\}$, of the set $\{(l,r) \in \mathcal{V};\ l < t < r\}$ such that $\mathcal{C}_m$ satisfies:

$$
\begin{aligned}
&(1)\ \mathcal{C}_m = \{D_i;\ i = 1, ..., m_t\},\ \text{ where } m_t\ (\le m),\\
&\qquad D_i = \Big\{(l,r) \in \mathcal{V}:\ \begin{cases} l \in (l_i - a_i, l_i], r \in [r_i, r_i + b_i) & \text{if } a_i, b_i > 0 \\ l = l_i\ (\text{or } r = r_i) & \text{if } a_i = 0,\ (\text{or } b_i = 0) \end{cases}\Big\},\\
&(2)\ l_1 - a_1 \le l_1 \le \cdots \le l_{m_t} - a_{m_t} \le l_{m_t} \le t \le r_{m_t} \le r_{m_t} + b_{m_t} \le \cdots \le r_1 \le r_1 + b_1,\\
&(3)\ l_{m_t} = \sup\{l;\ (l,r) \in \mathcal{V}, l < t < r,\ \mathrm{P}\{X \in (l,t]\} > 0\},\\
&(4)\ r_{m_t} = \inf\{r;\ (l,r) \in \mathcal{V}, l < t < r,\ \mathrm{P}\{X \in (l,t]\} > 0\},\\
&(5)\ G(l_i-, +\infty) - G(l_i - a_i, +\infty) \le 4/m,\ F(l_i-) - F(l_i - a_i) \le 4/m,\\
&(6)\ G(+\infty, r_i + b_i-) - G(+\infty, r_i) \le 4/m,\ F(r_i + b_i-) - F(r_i) \le 4/m,\ 1 \le i \le m_t,\\
&(7)\ \mathrm{P}\{(L,R) \in \cup_{i=1}^{m_t} D_i\} = \mathrm{P}\{L < t < R\},
\end{aligned}
\tag{4.4}
$$

where $l_0 = -\infty$, and $r_0 = \infty$. WLOG, we can assume that $a_i = l_i - l_{i-1}$ and $b_i = r_{i-1} - r_i$, $i = 1, ..., m_t$. Then D_i, (5) and (6) in (4.4) reduce to

$$
\begin{aligned}
&D_i = \{Y;\ L \in (l_{i-1}, l_i], R \in [r_i, r_{i-1})\},\\
&(5')\ G(l_i-, +\infty) - G(l_{i-1}, +\infty) \le 4/m,\ F(l_i-) - F(l_{i-1}) \le 4/m,\\
&(6')\ G(+\infty, r_{i-1}-) - G(+\infty, r_i) \le 4/m,\ F(r_{i-1}-) - F(r_i) \le 4/m,\ 1 \le i \le m_t.
\end{aligned}
$$

In the proof of (4.2), we need a condition: $t \in A_m$, where

$A_m = \{t \in \mathcal{B};\ \mathrm{P}\{X \in [l_{m_t}, r_{m_t}]\} \ge 1/\sqrt{m}\}$.

However, it can be shown (see Lemma 6.3) that

(S3) if (4.2) and (4.3) hold for all $t \in A_m$, then (4.2) and (4.3) hold for $t \in \mathcal{B} \setminus A_m$.

Thus, WLOG, we can assume that $t \in A_m$. It follows from (2) in (4.4) that

$$
\mathrm{P}\{X \in [l_i, r_i]\} \ge \mathrm{P}\{X \in [l_{m_t}, r_{m_t}]\} > 1/\sqrt{m}\ \text{ for } i = 1, ..., m_t. \tag{4.5}
$$

To establish an expression for $Q_n(t)$ corresponding to (3.7), we further denote

$$
\begin{aligned}
&Y = \{L, R\},\ Y^* = \{L^*, R^*\},\ y_i = \{l_i, r_i\},\\
&D_i^- = \{Y;\ L \in (l_{i-1}, l_i]\ \&\ R \in (r_i, r_{i-1});\ \text{or } L \in (l_{i-1}, l_i)\ \&\ R \in [r_i, r_{i-1})\},\\
&N_{i*-} = \#\{\text{CO's in } [l_i, r_i]\},\quad N_{i*-}^t = \#\{\text{CO's in } [l_i, t]\},\quad \beta_{i*-}(t) = 1[N_{i*-}^t > 0],
\end{aligned}
$$

for all possible i. Given $t \in [a,b]$, for $i = 1, ..., m_t$, let

$q_i = \#\{j; Y_j^* \in D_i\}$, $q_i^- = \#\{j;\ Y_j^* \in D_i^-\}$ $(q_i^- = q_i - q_i^0)$,

$q_i^+ = \#\{j;\ L_j^* = R_j^* = l_i \text{ or } r_i\}$, $\triangle_i = \#\{\text{CO's in } (l_{i-1}, l_i) \text{ or in } (r_i, r_{i-1})\}$.

Using the same idea as in proving (3.7), it can be shown that

$$
\begin{aligned}
&\sum_{h=1}^{m_t} \Big\{ \frac{q_h^-}{n} \prod_{1 \le i < h} \Big(\frac{N_{i*-} + q_i^- + \triangle_i}{N_{i*-} + \triangle_i}\Big)^{\beta_{i-*}(t)} \prod_{1 \le i < h} \Big(\frac{N_{i*} + q_i^0 + \triangle_i}{N_{i*} + \triangle_i}\Big)^{\beta_{i*}(t)} \frac{N_{h*-}^t}{(N_{h*-})^{\beta_{h*-}(t)} + \triangle_h}\\
&+ \frac{q_h^0}{n} \prod_{1 \le i \le h} \Big(\frac{N_{i*-} + q_i^- + \triangle_i}{N_{i*-} + \triangle_i}\Big)^{\beta_{i-*}(t)} \prod_{1 \le i < h} \Big(\frac{N_{i*} + q_i^0 + \triangle_i}{N_{i*} + \triangle_i}\Big)^{\beta_{i*}(t)} \frac{N_{h*}^t}{(N_{h*})^{\beta_{h*}(t)}} \Big\} \le Q_n(t)
\end{aligned}
\tag{4.6}
$$

and

$$Q_n(t) \le \sum_{h=1}^{m_t} \Big\{ \frac{q_h^-}{n} \prod_{1\le i<h} \Big(\frac{N_{i*-}+q_i^-}{N_{i*-}}\Big)^{\beta_{i-*}(t)} \prod_{1\le i<h} \Big(\frac{N_{i*}+q_i^0}{N_{i*}}\Big)^{\beta_{i*}(t)} \frac{N_{h*-}^t+\triangle_h}{(N_{h*-})^{\beta_{h*-}(t)}+\triangle_h}$$
$$+ \frac{q_h^0}{n} \prod_{1\le i\le h} \Big(\frac{N_{i*-}+q_i^-}{N_{i*-}}\Big)^{\beta_{i-*}(t)} \prod_{1\le i<h} \Big(\frac{N_{i*}+q_i^0}{N_{i*}}\Big)^{\beta_{i*}(t)} \frac{N_{h*}^t}{(N_{h*})^{\beta_{h*}(t)}} \Big\}. \quad (4.7)$$

The proof is relegated to the Appendix (see Lemma 6.4). We will use (4.7) to prove (4.2) and use (4.6) to prove (4.3).

We first show (4.2). By reasoning similar to that before equation (3.12), WLOG, we can assume $\beta_{h*}(t) = \beta_{h*-}(t) = 1$ and thus we omit the exponents $\beta_{h*}(t)$ etc. in the expressions of (4.7). Note that if G is discrete and only takes values at (l_i, r_i), $i = 1, ..., m_t$, all the ratio indexed with "$*-$" vanish and the expression on the right hand side of (4.7) reduces to (3.7). On the other hand, if G is continuous, then $q_h^0 = 0$ and $q_h = q_h^-$ w.p.1, and (4.7) reduces to

$$Q_n(t) \le \sum_{h=1}^{m_t} \Big\{ \Big(\frac{q_h}{n}\Big) \Big(\prod_{1\le i<h} \frac{N_{i*-}+q_i}{N_{i*-}} \Big) \Big(\frac{N_{h*-}^t+\triangle_h}{N_{h*-}+\triangle_h} \Big) \Big\}. \quad (4.8)$$

If G is neither continuous nor discrete, the proof is similar to that for the continuous case. For ease in understanding, we will assume that the joint distribution function $G(l, r)$ is continuous. Thus we will prove (4.8), instead of (4.7).

We now derive the limits for the three factors in the summation of (4.8). The following argument is parallel to (3.9) through (3.16) in the proof of Theorem 3.1. The limit of q_h/n is given by $\lim_{n\to\infty} \frac{q_h}{n} = \mathrm{P}\{Y^* \in D_h\}$ a.s. (see (3.9)) and

$$\begin{aligned} \mathrm{P}\{Y^* \in D_h\} &= \mathrm{P}\{X \in [l_h, r_h]\}\mathrm{P}\{Y \in D_h\} + \mathrm{P}\{X \in (L, l_h) \cup (r_h, R), Y \in D_h\} \\ &= \mathrm{P}\{X \in [l_h, r_h]\}\mathrm{P}\{Y \in D_h\} + O(1/m^2). \end{aligned} \quad (4.9)$$

Thus the limit of the first factor in (4.8) is

$$\lim_{n\to\infty} \frac{q_h}{n} = \mathrm{P}\{X \in [l_h, r_h]\}\mathrm{P}\{Y \in D_h\} + O(1/m^2) \text{ a.s..} \quad (4.10)$$

Since G is continuous, $\mathrm{P}\{(L, R) \supset [l, r]\}$ =$\mathrm{P}\{(L, R) \supset (l, r)\}$. Then

$$\begin{aligned} &\lim_{n\to\infty} \frac{N_{i*-}+q_i}{N_{i*-}} \quad (= \lim_{n\to\infty} (1 + \frac{q_i/n}{N_{i*-}/n})) \\ =&1 + \frac{\mathrm{P}\{Y^* \in D_i\}}{\mathrm{P}\{\text{CO's in } [l_i, r_i]\}} \\ =&1 + \frac{\mathrm{P}\{X \in [l_i, r_i]\}\mathrm{P}\{Y \in D_i\} + O(1/m^2)}{\mathrm{P}\{X \in [l_i, r_i]\} - \mathrm{P}\{X \in [l_i, r_i], (L, R) \supset [l_i, r_i]\}} \quad \text{(see (4.9))} \\ =&\frac{\mathrm{P}\{X \in [l_i, r_i]\}[1 - \sum_{j\le i} \mathrm{P}\{Y \in D_j\} + \mathrm{P}\{Y \in D_i\}] + O(1/m^2)}{\mathrm{P}\{X \in [l_i, r_i]\}[1 - \sum_{j\le i} \mathrm{P}\{Y \in D_j\}]} \quad (4.11) \\ =&\frac{[1 - \sum_{j<i} \mathrm{P}\{Y \in D_j\}] + O(m^{-3/2})}{[1 - \sum_{j\le i} \mathrm{P}\{Y \in D_j\}]} \text{ a.s. (due to (4.5)).} \end{aligned}$$

As a consequence, the second factor in (4.8) satisfies:

$$\begin{aligned}
&\overline{\lim_{n\to\infty}} \prod_{1\le i<h} \left(\frac{N_{i*-}+q_i}{N_{i*-}}\right) \\
\le &\prod_{1\le i<h} \left[\frac{1-\sum_{j<i} \mathrm{P}\{Y\in D_j\}+O(m^{-3/2})}{1-\sum_{j\le i} \mathrm{P}\{Y\in D_j\}}\right] \\
\le &\frac{1+O(m^{-3/2})}{1-\sum_{j\le h-1} \mathrm{P}\{Y\in D_j\}} \prod_{2\le i<h-1} \left[\frac{1-\sum_{j<i} \mathrm{P}\{Y\in D_j\}+O(m^{-3/2})}{1-\sum_{j<i} \mathrm{P}\{Y\in D_j\}}\right] \\
\le &\frac{1}{1-\sum_{j\le h-1} \mathrm{P}\{Y\in D_j\}} \left[\frac{1-\sum_{j<m_t} \mathrm{P}\{Y\in D_j\}+O(m^{-3/2})}{1-\sum_{j<m_t} \mathrm{P}\{Y\in D_j\}}\right]^{m_t}. \qquad (4.12)
\end{aligned}$$

Note that the summation in the denominator of the last expression

$$\sum_{j<m_t} \mathrm{P}\{Y\in D_j\} = \mathrm{P}\{(L,R)\supset (l_{m_t-1}, r_{m_t-1})\} \le \mathrm{P}\{L<t<R\} < 1 \ \text{ (see (2.2)).}$$

since $l_{m_t-1} < t < r_{m_t-1}$, $t\in\mathcal{O}$, which is an open set. Denote $d_0 = 1-\mathrm{P}\{L<t<R\}$, which is independent of the integer m and is >0. It follows from (4.12) that the second factor in (4.8) satisfies:

$$\begin{aligned}
\overline{\lim_{n\to\infty}} \prod_{1\le i<h} \left(\frac{N_{i*-}+q_i}{N_{i*-}}\right) \le &\frac{1}{1-\sum_{j\le h-1} \mathrm{P}\{Y\in D_j\}} \left[\frac{d_0+O(m^{-3/2})}{d_0}\right]^{m_t} \\
\le &\frac{1}{1-\sum_{j\le h-1} \mathrm{P}\{Y\in D_j\}} \left[1+O(m^{-1/2}/d_0)\right] \text{ a.s..} \qquad (4.13)
\end{aligned}$$

Note that by the assumption on $\mathcal{C}_m$ we have

$$\begin{aligned}
\lim_{n\to\infty} \left|\frac{\triangle_h}{n}\right| =&\mathrm{P}\{X\in (l_{h-1},l_h)\cup(r_h,r_{h-1})\} - \mathrm{P}\{X\in (l_{h-1},l_h)\cup(r_h,r_{h-1}), (L^*,R^*)\supset (l_h,r_h)\} \\
\le &\mathrm{P}\{X\in (l_{h-1},l_h)\cup(r_h,r_{h-1})\}[1-\mathrm{P}\{(L,R)\supset (l_{h-1},r_{h-1})\}] \\
\le &(8/m)d_0 \quad \text{(by (5') and (6') in (4.4) and by (4.12)).}
\end{aligned}$$

Hence, the last ratio in (4.8) satisfies

$$\begin{aligned}
&\lim_{n\to\infty} \frac{N^t_{h*-}+\triangle_h}{N_{h*-}+\triangle_h} \le \lim_{n\to\infty} \frac{N^t_{h*-}+O(1/m))}{N_{h*-}} \\
=&\frac{[\mathrm{P}\{X\in[l_h,t]\} - \mathrm{P}\{X\in[l_h,t], L^*<t<R^*\}]+O(1/m)}{\mathrm{P}\{X\in[l_h,r_h]\} - \mathrm{P}\{X\in[l_h,r_h], (L,R)\supset[l_h,r_h]\}} \\
=&\frac{\mathrm{P}\{X\in[l_h,t]\} - \sum_{i=1}^h \mathrm{P}\{X\in[l_h,t], (L,R)\in D_i\}}{\mathrm{P}\{X\in[l_h,r_h]\}[1-\sum_{i=1}^h \mathrm{P}\{Y\in D_i\}]}
\end{aligned}$$

$$+\frac{-\sum_{i>h}^{m_t}\mathrm{P}\{X\in(L,t],Y\in D_i\}}{\mathrm{P}\{X\in[l_h,r_h]\}[1-\sum_{i=1}^{h}\mathrm{P}\{Y\in D_i\}]}+O(\frac{1}{d_0\sqrt{m}})\quad(\text{due to (4.5)})$$

$$=\frac{\mathrm{P}\{X\in[l_h,t]\}-\sum_{i=1}^{h}\mathrm{P}\{X\in[l_h,t]\}\mathrm{P}\{Y\in D_i\}}{\mathrm{P}\{X\in[l_h,r_h]\}[1-\sum_{i=1}^{h}\mathrm{P}\{Y\in D_i\}]}$$

$$+\frac{-\sum_{i>h}^{m_t}\mathrm{P}\{X\in[l_i,t]\}\mathrm{P}\{Y\in D_i\}+O(1/m)[1-\sum_{i=1}^{h}\mathrm{P}\{Y\in D_i\}]}{\mathrm{P}\{X\in[l_h,r_h]\}[1-\sum_{i=1}^{h}\mathrm{P}\{Y\in D_i\}]}+O(\frac{1}{d_0\sqrt{m}})$$

$$=\frac{\mathrm{P}\{X\in[l_h,t]\}}{\mathrm{P}\{X\in[l_h,r_h]\}}-\frac{\sum_{i>h}^{m_t}\mathrm{P}\{X\in[l_i,t]\}\mathrm{P}\{Y\in D_i\}}{\mathrm{P}\{X\in[l_h,r_h]\}[1-\sum_{i=1}^{h}\mathrm{P}\{Y\in D_i\}]}+O(\frac{1}{d_0\sqrt{m}})\ \text{a.s..}\qquad(4.14)$$

(4.10), (4.13) and (4.14) give the limits of the three factors in (4.8), thus

$$\begin{aligned}\overline{\lim_{n\to\infty}}\,Q_n(t)\le&\sum_{h=1}^{m_t}\Big\{[\mathrm{P}\{X\in[l_h,r_h]\}\mathrm{P}\{Y\in D_h\}+O(1/m^2)]\frac{1+O(m^{-1/2})}{1-\sum_{j<h}\mathrm{P}\{Y\in D_j\}}\\&\cdot[\frac{\mathrm{P}\{X\in[l_h,t]\}}{\mathrm{P}\{X\in[l_h,r_h]\}}-\frac{\sum_{i>h}^{m_t}\mathrm{P}\{X\in[l_i,t]\}\mathrm{P}\{Y\in D_i\}}{\mathrm{P}\{X\in[l_h,r_h]\}[1-\sum_{i=1}^{h}\mathrm{P}\{Y\in D_i\}]}+O(1/(\sqrt{m})]\Big\}\\\le&\sum_{h=1}^{m_t}\Big\{\frac{\mathrm{P}\{X\in[l_h,t]\}\mathrm{P}\{Y\in D_h\}}{1-\sum_{j<h}\mathrm{P}\{Y\in D_j\}}-\frac{\mathrm{P}\{Y\in D_h\}\sum_{i>h}^{m_t}\mathrm{P}\{X\in[l_i,t]\}\mathrm{P}\{Y\in D_i\}}{[1-\sum_{j<h}\mathrm{P}\{Y\in D_j\}][1-\sum_{i=1}^{h}\mathrm{P}\{Y\in D_i\}]}\Big\}\\&+O(m^{-1/2})\ \text{a.s..}\qquad(4.15)\end{aligned}$$

Moreover

$$\begin{aligned}\mathrm{P}\{L<X\le t<R\}=&\sum_{h=1}^{m_t}\mathrm{P}\{L<X\le t,Y\in D_h\}\\=&\sum_{h=1}^{m_t}[\mathrm{P}\{l_h\le X\le t\}+O(1/m)]\mathrm{P}\{Y\in D_h\}\\=&\sum_{h=1}^{m_t}\mathrm{P}\{X\in[l_h,t]\}\mathrm{P}\{Y\in D_h\}+O(1/m).\qquad(4.16)\end{aligned}$$

Thus it follows from (4.15) and (4.16) that

$$\begin{aligned}&\overline{\lim_{n\to\infty}}\,\{Q_n(t)-\mathrm{P}\{L<X\le t<R\}\}\\\le&\sum_{h=1}^{m_t}\Big\{\frac{\mathrm{P}\{X\in[l_h,t]\}\mathrm{P}\{Y\in D_h\}}{1-\sum_{j<h}\mathrm{P}\{Y\in D_j\}}-\frac{\mathrm{P}\{Y\in D_h\}\sum_{i>h}^{m_t}\mathrm{P}\{X\in[l_i,t]\}\mathrm{P}\{Y\in D_i\}}{[1-\sum_{j<h}\mathrm{P}\{Y\in D_j\}][1-\sum_{i=1}^{h}\mathrm{P}\{Y\in D_i\}]}\Big\}\\&-\sum_{h=1}^{m_t}\mathrm{P}\{X\in[l_h,t]\}\mathrm{P}\{Y\in D_h\}+O(m^{-1/2})\ \text{a.s..}\qquad(4.17)\end{aligned}$$

Notice that expression (4.17) is identical to expression (3.15) except that m_t is replaced by m, $Y\in D_h$ by $(L,R)=(l_h,r_h)$, $[l_i,t]$ by $(l_i,t]$ and $O(1/m^{-1/2})$ by 0. Thus

$$\overline{\lim_{n\to\infty}}\,\{Q_n(t)-\mathrm{P}\{L<X\le t<R\}\}\le\sum_{h=1}^{m_t}s_h+O(m^{-1/2})\ \text{a.s.},$$

where s_h is the same as the summand in (3.16), except that $(L, R) = (l_h, r_h)$ is replaced by $Y \in D_h$ and $(l_i, t]$ is replaced by $[l_i, t]$. The same argument as in the proof of Lemma 6.2 yields $\sum_{h=1}^{m_t} s_h = 0$. (4.2) then follows.

With the same idea, we can show that (4.3) holds. This completes the proof of (S1).

In a similar manner, we can show (S2). This completes the proof. □

If $\tau_l = +\infty$, Theorem 4.1 yields (3.1). If $\tau_l < \infty$, in order to show (3.1), we need to further verify (S2) of (4.1) for $t = \tau_l$ and (S1) of (4.1) for $t = \tau_r$. It can be verified that the proof of Theorem 4.1 can be applied to all $t \in \overline{\mathcal{O}}$, provided that for each $t \in \overline{\mathcal{O}}$,

$$\mathrm{P}\{(L, R) \supset [l_{m_t}, r_{m_t}]\} \quad (= 1 - d_0) \quad < 1. \tag{4.18}$$

(4.18) is needed in (4.13) etc.. However, (4.18) may not hold on the boundary of $\mathcal{O}$.

To show (4.1) on the boundary of $\mathcal{O}$, we first establish three lemmas.

Lemma 4.2. *For any arbitrary F and G,*

$$\overline{\lim_{n\to\infty}}\ \hat{S}_I(\tau_l) \le \overline{\lim_{n\to\infty}}\ \hat{S}_I(\tau_l-) \le S(\tau_l-) \text{ a.s. and } \underline{\lim_{n\to\infty}}\ \hat{S}_I(\tau_r) = \underline{\lim_{n\to\infty}}\ \hat{S}_I(\tau_r+) \ge S(\tau_r) \text{ a.s..}$$

It is worth noting that due to our convention $\hat{S}_I(t)$ is right continuous, thus $\hat{S}_I(\tau_r) = \hat{S}_I(\tau_r+)$ in the lemma. The proof of the first inequality is identical to Yu and Li (1994, Lemma 2). The proof for the second inequality is similar to that for the first one.

Lemma 4.3. *For any arbitrary F and G,*
(1) if $F(\tau_l-) = 1$, then (S2) of (4.1) holds for $t = \tau_l$;
(2) if $F(\tau_r) = 0$, then (S1) of (4.1) holds for $t = \tau_r$.

The proof of the first inequality is the same as Yu and Li (1994, Lemma 4). The proof for the second inequality is similar to that for the first one.

Lemma 4.4. *For any F and G,*
(1) if $P\{L = \tau_l\} > 0$ then (S2) holds for $t = \tau_l$;
(2) if $P\{R = \tau_r\} > 0$ then (S1) holds for $t = \tau_r$.

Proof: We first prove statement (1). Note that if $t = \tau_l$ and $\mathrm{P}\{L = \tau_l\} > 0$, then $d_0 > 0$ (see (4.18)) and $\mathrm{P}\{(L, R) \supset (l_i, r_i)\} \le 1 - d_0 < 1$ for $t = \tau_l$. It can be checked that in the proof of Theorem 4.1 all the statements from (4.6) through the end of the proof hold for $t = \tau_l$. This completes the proof of Statement (1). Statement (2) can be proved similarly. □

Lemma 4.5. *For any arbitrary F and G, (S2) in (4.1) holds for $t = \tau_l$ and (S1) in (4.1) holds for $t = \tau_r$.*

Note that Lemmas (4.3) and (4.4) are special cases of the theorem. The proof of the theorem is relegated to the Appendix.

Remark 4.1. When $G(\tau_l-, +\infty) = 1$, if $t = \tau_l$ then we would never observe any exact observation at τ_l (w.p.1). Thus due to the convention in section 2, we have $\hat{S}_I(\tau_l-) = \hat{S}_I(\tau_l) = \hat{S}_I(L^*_{(n)})$. In particular, if $G(t, +\infty)$ is continuous at $t = \tau_l$ but $F(t)$ is not, then

$$\lim_{n\to\infty} \hat{S}_I(\tau_l) = \lim_{n\to\infty} \hat{S}_I(\tau_l-) = S(\tau_l-) > S(\tau_l) \text{ a.s..}$$

However, due to the convention on $\hat{S}_I(t)$, $\hat{S}_I(t)$ is right continuous at τ_r and so is $S(t)$. Thus, $\hat{S}_I(\tau_r)$ always converges to $S(\tau_r)$ a.s..

In view of Remark 4.1, it is easy to derive the following result:

Lemma 4.6. *If $F(t)$ is continuous at τ_l, then* $\lim_{n\to\infty} \sup_{t\in\overline{\mathcal{O}}} |\hat{S}_I(t) - S(t)| = 0$ a.s..

It follows from Lemmas 4.1, 4.3, 4.4, 4.5 and 4.6 that the following result holds.

Theorem 4.2. *Let $\mathcal{O}^* = \mathcal{O} \cup \{\tau_r\}$. Under the DI Model, for any arbitrary F and G,*

$$\lim_{n\to\infty} \sup_{t\in\mathcal{O}^*} |\hat{S}_I(t) - S(t)| = 0 \text{ a.s.;}$$

$$\lim_{n\to\infty} \sup_{t\in\overline{\mathcal{O}}} |\hat{S}_I(t) - S(t)| = 0 \text{ a.s. } \textit{unless } F(\tau_l-) < F(\tau_l) \textit{ and } G(\tau_l-, +\infty) = 1.$$

5. Discussion. A direct consequence of Theorem 4.2 is the following result related to the PLE with left-censored data.

Corollary 1. *For any arbitrary F and G, the PLE $\hat{S}_{PL}(t)$ with left-censored data satisfies* $\lim_{n\to\infty} \sup_{t\geq\tau_r} |\hat{S}_{PL}(t) - S(t)| = 0$ a.s..

It is interesting to see that the PLE with left-censored data does not carry over the short-coming of the PLE with right-censored data at τ_l. As implied by Theorem 4.2,

$$\lim_{n\to\infty} \sup_{t\leq\tau_l} |\hat{S}_{PL}(t) - S(t)| = 0 \text{ a.s.}$$

failed for arbitrary F and G with right-censored data. The short-coming does not depend on the definition of $\hat{S}_I(t)$ for $t > L^*_{(n)}$.

With right-censored data, it has been proved that

$$\text{For any arbitrary } F \text{ and } G, \lim_{n\to\infty} \sup_{t\leq L^*_{(n)}} |\hat{S}_c(t) - S(t)| = 0 \text{ a.s.}$$

(see Yu and Li (1994)). The corresponding statements with interval-censored data are

$$\lim_{n\to\infty} \sup_{\min_i R^*_i \leq t, \text{ or } t\leq L^*_{(n)}} |\hat{S}_c(t) - S(t)| = 0 \text{ a.s.,} \tag{5.1}$$

which follow from Theorem 4.2.

Finally, it is desirable to derive the almost sure limit of $\hat{S}_I(t)$ over the entire region $\{t \geq 0\}$. Define $M = \begin{cases} 0 & \text{if } \tau_l = 0 \text{ and } \tau_r < \infty \\ \frac{\tau_l+\tau_r}{2} & \text{if } 0 < \tau_l \text{ and } \tau_r < +\infty \\ +\infty & \text{if } 0 < \tau_l \text{ and } \tau_r = +\infty. \end{cases}$ It is easy to derive the following result.

Remark 5.1. $\lim_{n\to\infty} \sup_{M\leq t\leq\tau_r} |\hat{S}_c(t) - S(\tau_r)| = 0$ a.s. and

$$\begin{cases} \lim_{n\to\infty} \sup_{\tau_l\leq t<M} |\hat{S}_c(t) - S(\tau_l-)| = 0 & \text{if } S(\tau_l-) > S(\tau_l) \text{ and } G(\tau_l-, +\infty) = 1; \\ \lim_{n\to\infty} \sup_{\tau_l\leq t<M} |\hat{S}_c(t) - S(\tau_l)| = 0 & \text{otherwise} \end{cases} \text{ a.s..}$$

It is well known that the GMLE is not uniquely defined in empty intervals. However the uniformly strong consistency of the GMLE does depend on the definition over empty intervals (see Yu and Li (1994)).

The natural extension of $\hat{S}_I(t)$ from expression (2.1) assigns weight only to R_i^*. However, this convention will affect both (3.1) and Theorem 4.2. Finally, we point out that (5.1) does not depend on the definition of $\hat{S}_I(t)$ over empty intervals.

Reference.

* Chang, M.N. and Yang, G. (1987). Strong consistency of a self-consistent estimator of the survival function with doubly censored data. *Ann. Statist.* 15, 1536-1547.
* Efron, B (1967). The two sample problem with censored data. *Fifth Berkeley Symposium on Mathematical Statistics.* University of California Press, 831-853.
* Gill, R. (1983). Large sample behavior of the product-limit estimator on the whole line. *Ann. Statist.* 11, 49-58.
* Gomez, G. Julia, O. and Utzet, P. (1992). Survival analysis for left censored data. In *Survival Analysis: State of the Art, Klein, J. P. and Goel, P. K. Ed.*, Kluwer, Netherlands.
* Groeneboom, P. and Wellner, J. A. (1992). Information bounds and nonparametric maximum likelihood estimation. *Birkhäuser Verlag*, Basel.
* Gu, M.G. and Zhang, C-H. (1993) Asymptotic properties of self-consistent estimator based on doubly censored data. *Ann. Statist.* 21. 611-624.
* Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *J. Amer. Stat. Assoc.*, 53, 457-481.
* Kiefer, J and Wolfowitz, J. (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *Ann. Math. Statist.* 27, 887-906.
* Peto, R. (1973). Experimental survival curves for interval-censored data. *Appl. Statist.* 22, 86-91.
* Stute, W and Wang, J.-L. (1993) The strong law under random censorship. *Ann. Statist.* (to appear).
* Tsai, W. and Crowley, J. (1985). A large sample study of the generalized maximum likelihood estimators from incomplete data via self-consistency. *Ann. Statist.* 13, 1317-1334.
* Turnbull, B. W. (1976). The empirical distribution function with arbitrary grouped, censored and truncated data. *J. Roy. Statist. Soc. Ser.* B, 38, 290-295.
* Yu, Qiqing and Li, Linxiong (1994). On the strong uniform consistency of the product limit estimator. *Sankhyā. A.* 56.
* Yu, Qiqing and Wong, George (1993). Estimation of a survival function with interval-censored data under the DI Model. Submitted for publication.

Appendix

We give the proofs of lemmas in this section. This section could be deleted and put in a technical report in a future revision.

Proof of Lemma 4.1: Since F is bounded, monotone and right continuous on $[a,b)$, for any $\epsilon > 0$, there exist finitely many points, $t_1, ..., t_k \in [a,b)$ such that $a = t_1 < t_2 < \cdots < t_k$, $|F(t_{i+1}-) - F(t_i)| < \epsilon$, for $i = 1, ..., k$, where $t_{k+1} = b$. Since $F_n(t)$ converges to $F(t)$ for all $t = t_i$, $i = 1, ..., k$, and for all $t = t_i-$, $i = 2, ..., k+1$, there exists an n_0 such that whenever $n \geq n_0$ we have $|F_n(t) - F(t)| < \epsilon$, for all $t = t_i$, $i = 1, ..., k$, and for all $t = t_i-$, $i = 2, ..., k+1$. For any $t \in [a,b)$, there exists some i such that $t \in [t_i, t_{i+1})$, then

$$
\begin{aligned}
&|F_n(t) - F(t)| \qquad (6.1)\\
\leq &\max\{|F_n(t_i) - F(t_{i+1}-)|, |F_n(t_{i+1}-) - F(t_i)|\}\\
\leq &\max\{|F_n(t_i) - F(t_i) + F(t_i) - F(t_{i+1}-)|, |F_n(t_{i+1}-) - F(t_{i+1}-) + F(t_{i+1}-) - F(t_i)|\}\\
\leq &2\epsilon,
\end{aligned}
$$

whenever $n \geq n_0$. Since t and ϵ are arbitrary, the lemma follows. □

Proof of Lemma 4.5: WLOG, we can assume that $0 < \tau_l < \tau_r < +\infty$. It follows from Lemmas 4.1, 4.3 and 4.4 that it suffices to show that

$$\lim_{n\to\infty} |\hat{S}_I(\tau_l-) - S(\tau_l-)| = 0 \text{ a.s., if } F(\tau_l-) < 1 \text{ and } G(\tau_l-, +\infty) = 1; \qquad (6.2)$$

$$\lim_{n\to\infty} |\hat{S}_I(\tau_r+) - S(\tau_r)| = 0 \text{ a.s., if } F(\tau_r) > 0 \text{ and } G(+\infty, \tau_r) = 0. \qquad (6.3)$$

We first show (6.2). Notice that $\overline{\lim_{n\to\infty}} \hat{S}_I(\tau_l-) \leq S(\tau_l-)$ a.s. by Lemma 4.2. Then (2.1), (3.2) and (3.3) yield

$$\underline{\lim_{n\to\infty}} Q_n(\tau_l-) \geq \mathrm{P}\{L < X < \tau_l < R\}. \qquad (6.4)$$

Thus it suffices to show that

$$\overline{\lim_{n\to\infty}} Q_n(\tau_l-) \leq \mathrm{P}\{L < X < \tau_l < R\} \text{ if } F(\tau_l-) < 1 \text{ and } G(\tau_l-, +\infty) = 1. \qquad (6.5)$$

Hereafter, we assume $F(\tau_l-) < 1$, $G(\tau_l-, +\infty) = 1$ and $t = \tau_l-$ (see (6.2)), then it yields $l_{m_t} = \tau_l$ (see (4.4)) and $r_{m_t} = \tau_r$. Thus $d_0 = 1 - \mathrm{P}\{(L,R) \supset (l_{m_t}, r_{m_t})\} = 0$ and the proof of Theorem 4.1 is not applicable directly, since it needs $d_0 > 0$ (see (4.18) or (4.13)). To mimic the proof of Theorem 3.1 or 4.1, we modify (4.7) as follows:

$$
\begin{aligned}
&Q_n(t)\\
\leq &\sum_{h=1}^{m_t-1} \Big\{ \frac{q_h^-}{n} \prod_{1\leq i<h} \Big(\frac{N_{i*-} + q_i^-}{N_{i*-}}\Big) \prod_{1\leq i<h} \Big(\frac{N_{i*} + q_i^0}{N_{i*}}\Big) \frac{N^t_{h*-} + \triangle_h}{(N_{h*-})^{\beta_{h*-}(t)} + \triangle_h}\\
&+ \frac{q_h^0}{n} \prod_{1\leq i\leq h} \Big(\frac{N_{i*-} + q_i^-}{N_{i*-}}\Big) \prod_{1\leq i<h} \Big(\frac{N_{i*} + q_i^0}{N_{i*}}\Big) \frac{N^t_{h*}}{(N_{h*})^{\beta_{h*}(t)}} \Big\} \qquad (6.6)\\
&+ \frac{q_{m_t}^-}{n} \prod_{1\leq i<m_t} \Big(\frac{N_{i*-} + q_i^-}{N_{i*-}}\Big) \prod_{1\leq i<m_t} \Big(\frac{N_{i*} + q_i^0}{N_{i*}}\Big) \max\Big\{ \frac{N^t_{j-}}{(N_{j-})^{\beta_{j-}(t)}};\ L^*_{(j)} \in (l_{m_t-1}, \tau_l) \Big\}\\
&(+ \frac{q_{m_t}^0}{n} \prod_{1\leq i\leq m_t} \Big(\frac{N_{i*-} + q_i^-}{N_{i*-}}\Big) \prod_{1\leq i<m_t} \Big(\frac{N_{i*} + q_i^0}{N_{i*}}\Big) \frac{N^t_{m_t*}}{(N_{m_t*})^{\beta_{m_t*}(t)}} \quad \text{(which equals 0))}.
\end{aligned}
$$

(6.6) can be proved in a similar manner as in deriving (4.7). The only difference between (4.7) and (6.6) is in the third summand (indexed by m_t). Note that when $t = \tau_l-$, $N^t_{m_t*} = 0$ (see (3.8)) and thus the fourth expression in (6.6) equals 0.

WLOG, we can assume that $S(\tau_l-) - S(\tau_r) = a_0 > 0$. Otherwise, since $\tau_l < \tau_r$ and in view of Lemma 4.2, we have

$$\hat{S}_I(\tau_l-) \le S(\tau_l-) = S(\tau_r) \le \hat{S}_I(\tau_r) \le \hat{S}_I(\tau_l-) \text{ a.s..}$$

i.e., (6.2) and (6.3) hold.

Let the notation be the same as in the proof of Theorem 4.1, it can be verified that the arguments between (4.6) and (4.17) remain true except the following: $\mathrm{P}\{X \in [l_i, r_i]\} \ge a_0 > 0$ for $t = \tau_l-$, thus it does not need the restriction (4.5). In order to show (6.5) or equivalently (4.2), it suffices to show that inequalities (4.13) and (4.14) holds.

To this end, as argued in the paragraph after (4.8), WLOG, we can assume that G is continuous, and we further assume $\mathrm{P}\{Y \in D_i\} = 1/(m)$, $i = 1, ..., m_t$. (6.6) reduces to

$$Q_n(t) \le \sum_{h=1}^{m_t-1} \Big\{ \frac{q_h}{n} \prod_{1\le i<h} \Big(\frac{N_{i*-} + q_i}{N_{i*-}}\Big) \frac{N^t_{h*-} + \triangle_h}{(N_{h*-})^{\beta_{h*-}(t)} + \triangle_h} \tag{6.7}$$

$$+ \frac{q_{m_t}}{n} \prod_{1\le i<m_t} \Big(\frac{N_{i*-} + q_i}{N_{i*-}}\Big) \max\Big\{ \frac{N^t_{j-}}{(N_{j-})^{\beta_{j-}(t)}};\ L^*_{(j)} \in (l_{m_t-1}, \tau_l) \Big\}. \tag{6.8}$$

Since $\mathrm{P}\{X \in [l_i, r_i]\} \ge a_0 > 0$ for all i if $l_i < \tau_l < r_i$, in a similar manner as in deriving (4.13), we can show that for $t = \tau_l-$,

$$\begin{aligned}
&\lim_{n\to\infty} \prod_{1\le i<h} \Big(\frac{N_{i*-} + q_i}{N_{i*-}}\Big) \\
\le &\prod_{1\le i<h} \Big[\frac{1 - \sum_{j<i} \mathrm{P}\{Y \in D_j\} + O(1/m^2)}{1 - \sum_{j\le i} \mathrm{P}\{Y \in D_j\}}\Big] \\
\le &\frac{1 + O(1/m^2)}{1 - \sum_{j<h} \mathrm{P}\{Y \in D_j\}} \prod_{2\le i<h-1} \Big[\frac{1 - \sum_{j<i} \frac{1}{m} + O(1/m^2)}{1 - \sum_{j\le i-1} \frac{1}{m}}\Big] \\
\le &\frac{1}{1 - \sum_{j<h} \mathrm{P}\{Y \in D_j\}} \prod_{i=1}^{h-2} \frac{1 - \frac{i}{m} + O(1/m^2)}{1 - \frac{i}{m}} \quad \text{a.s.,} \quad h \le m_t.
\end{aligned}$$

(It is worth noting that $O(1/m^2)$ in the above expression is due to $\mathrm{P}\{X \in [l_i, r_i]\} \ge a_0 > 0$ and $O(1/m^{3/2})$ in (4.13) is due to $\mathrm{P}\{X \in [l_i, r_i]\} \ge 1/(m^{1/2})\}$.) Since

$$\begin{aligned}
1 \le \prod_{i=2}^{m-1} \frac{\frac{i}{m} + O(1/m^2)}{\frac{i}{m}} \le &\prod_{2\le i<\sqrt{m}} \frac{\frac{2}{m} + O(1/m^2)}{\frac{2}{m}} \cdot \prod_{i\ge\sqrt{m}}^{m-1} \frac{\frac{\sqrt{m}}{m} + O(1/m^2)}{\frac{\sqrt{m}}{m}} \\
\le &\Big(1 + O(1/m)\Big)^{\sqrt{m}} \cdot \Big(1 + O(1/(m^{3/2}))\Big)^{m-\sqrt{m}} \\
\le &1 + O(1/m^{1/2}),
\end{aligned} \tag{6.9}$$

$$\lim_{n\to\infty}\prod_{1\le i<h}\Big(\frac{N_{i*-}+q_i}{N_{i*-}}\Big)=\frac{1}{1-\sum_{j<h}\mathrm{P}\{Y\in D_j\}}(1+O(m^{-1/2})).$$

This is inequality (4.13) for $t=\tau_l-$.

It can be verified that the argument in proving (4.14) holds for $h=1,...,m_t-1$, which is the upper bound for the limit of the third factor in expression (6.7). We need to show a similar inequality corresponding to (4.14) holds for $h=m_t$. The third factor in expression (6.8) satisfies

$$\begin{aligned}
&\lim_{n\to\infty}\sup\Big\{\frac{N^t_{j-}}{N^{\beta_{j-}(t)}_{j-}};\ L^*_{(j)}\in(l_{m_t-1},\tau_l)\Big\}\\
\le&\sup\Big\{\frac{[\mathrm{P}\{X\in[l,t]\}-\mathrm{P}\{X\in[l,t],L^*<t<R^*\}]}{\mathrm{P}\{X\in[l,r]\}-\mathrm{P}\{X\in[l,r],(L^*,R^*)\supset[l,r]\}};\ l\in(l_{m_t-1},\tau_l),r\in(\tau_r,r_{m_t-1}),(l,r)\in\mathcal{V}\Big\}\\
\le&\sup\Big\{\frac{[\mathrm{P}\{X\in[l,t]\}-\mathrm{P}\{X\in[l,t],(L,R)\supset[l,r]\}]}{\mathrm{P}\{X\in[l,r]\}-\mathrm{P}\{X\in[l,r],(L,R)\supset[l,r]\}};\ l\in(l_{m_t-1},\tau_l),r\in(\tau_r,r_{m_t-1}),(l,r)\in\mathcal{V}\Big\}\\
\le&\sup\Big\{\frac{\mathrm{P}\{X\in[l,t]\}}{\mathrm{P}\{X\in[l,r]\}};\ l\in(l_{m_t-1},\tau_l),r\in(\tau_r,r_{m_t-1}),(l,r)\in\mathcal{V}\Big\}\qquad(6.10)\\
\le&O(1/(ma_0))\qquad\text{(due to (5$'$) in (4.4))}\\
=&\frac{\mathrm{P}\{X\in[\tau_l,t]\}}{\mathrm{P}\{X\in[\tau_l,\tau_r]\}}-\frac{\sum_{i>m_t}^{m_t}\mathrm{P}\{X\in[l_i,t]\}\mathrm{P}\{Y\in D_i\}}{\mathrm{P}\{X\in[\tau_l,\tau_r]\}[1-\sum_{i=1}^{m_t}\mathrm{P}\{Y\in D_i\}]}+O(1/m)\text{ a.s. (since }t=\tau_l-).
\end{aligned}$$

This is the inequality corresponding to (4.14) for $h=m_t$. Thus the arguments between (4.13) and (4.17) hold. As a consequence (6.5) holds. It follows from (6.4), (6.5) and (3.4) that (6.2) holds.

(6.3) can be shown using the same idea. This completes the proof of the lemma. □

Lemma 6.1. *Using the notation as in the proof of Theorem 3.1, (3.7) holds.*

Proof. We will prove the lemma by induction on m (see (3.7)).

$m=1$. By the definition of m, there is only one (l_1,r_1) in $\mathcal{V}$ such that $t\in(l_1,r_1)$. Suppose that there are q_1 ties, say, $(L^*_1,R^*_1)=\cdots=(L^*_{q_1},R^*_{q_1})=(l_1,r_1)$. It follows from the convention on the ties (see Remark 2.2) that

1. $N_j=N_{1*}+q_1-j$, $j\le q_1$, since $\{L^*_i,R^*_i\}$ is a CO of (L^*_{i-1},R^*_{i-1}) for $i=2,...,q_1$;
2. $N^t_1=\cdots=N^t_{q_1}=N^t_{1*}$, since $L^*_i<t<R^*_i$, $i\le q_1$;
3. $N^t_i=\beta_i(t)=0$ for $i=q_1+1,...,n$.

If $N^t_{1*}>0$ (and thus $\beta_i(t)=1$, $i\le q_1$), then (3.3) yields

$$\begin{aligned}
Q_n(t)=&\sum_{j=1}^{q_1}\frac{1}{n}[\prod_{1\le k<j}(1+\frac{1}{N_k})^{\beta_k(t)}]\frac{N^t_j}{(N_j)^{\beta_j(t)}}\\
=&\sum_{j=1}^{q_1}\frac{1}{n}[\prod_{1\le k<j}(1+\frac{1}{N_{1*}+q_1-k})]\frac{N^t_{1*}}{N_{1*}+q_1-j}\\
=&\frac{q_1}{n}[\frac{N^t_{1*}}{N_{1*}}]\quad\text{(which can be proved by induction on }q_1)\qquad(6.11)\\
=&[\frac{q_1}{n}][\prod_{1\le k<1}\frac{N_{k*}+q_k}{N_{k*}}][\frac{N^t_{1*}}{N_{1*}}].
\end{aligned}$$

That is, (3.7) holds if $N^t_{1*} > 0$. If $N^t_{1*} = 0$, (3.7) is trivially true. Thus (3.7) holds for $m = 1$.

Now assume that (3.7) holds for $m-1$, we will show that (3.7) holds for m. By assumption, there are m distinct (l_j, r_j)'s satisfying $l_1 \le \cdots \le l_m < t < r_m \le \cdots \le r_1$. WLOG, let $\{L^*_1, R^*_1\} = \cdots = \{L^*_{q_1}, R^*_{q_1}\} = \{l_1, r_1\}$, ..., $\{L^*_{q_1+\cdots+q_{m-1}+1}, R^*_{q_1+\cdots+q_{m-1}+1}\} = \cdots = \{L^*_{q_1+\cdots+q_m}, R^*_{q_1+\cdots+q_m}\} = \{l_m, r_m\}$, where $q_1 + \cdots + q_m \le n$. Then

$$
\begin{aligned}
Q_n(t) &= \sum_{j=1}^{q_1+\cdots+q_m} \frac{1}{n}\Big[\prod_{1\le k<j}(1+\frac{1}{N_k})^{\beta_k(t)}\Big]\frac{N^t_j}{(N_j)^{\beta_j(t)}} \\
&= \sum_{j=1}^{m-1}\Big\{[\frac{q_j}{n}][\prod_{1\le k<j}(\frac{N_{k*}+q_k}{N_{k*}})^{\beta_{k*}(t)}][\frac{N^t_{j*}}{(N_{j*})^{\beta_{j*}(t)}}]\Big\} \quad \text{(by induction assumption on } m-1) \\
&\quad + \sum_{j=q_1+\cdots+q_{m-1}+1}^{q_1+\cdots+q_m} \frac{1}{n}\Big[\prod_{1\le k<j}(1+\frac{1}{N_k})^{\beta_j(t)}\Big]\frac{N^t_j}{(N_j)^{\beta_j(t)}}. \qquad (6.12)
\end{aligned}
$$

WLOG, we can assume that $N^t_{j*} > 0$ (and thus $\beta_{j*}(t) = 1$), then the last summation equals

$$
\begin{aligned}
&\sum_{j=q_1+\cdots+q_{m-1}+1}^{q_1+\cdots+q_m} \frac{1}{n}\Big[\prod_{1\le k<j}(1+\frac{1}{N_k})^{\beta_j(t)}\Big]\frac{N^t_j}{(N_j)^{\beta_j(t)}} \\
&= \sum_{j=q_1+\cdots+q_{m-1}+1}^{q_1+\cdots+q_m} \frac{1}{n}\Big[\prod_{1\le k<m}\Big[\prod_{1\le s\le q_k}(1+\frac{1}{N_{k*}+q_k-s})\Big]\Big]\Big[\prod_{q_1+\cdots+q_{m-1}+1\le k<j}(1+\frac{1}{N_k})\Big]\frac{N^t_j}{N_j} \\
&= \Big[\prod_{1\le k<m}(\frac{N_{k*}+q_k}{N_{k*}})\Big]\sum_{j=q_1+\cdots+q_{m-1}+1}^{q_1+\cdots+q_m}\frac{1}{n}\Big[\prod_{q_1+\cdots+q_{m-1}+1\le k<j}(1+\frac{1}{N_k})\Big]\frac{N^t_j}{N_j} \qquad (6.13) \\
&= \Big[\prod_{1\le k<m}(\frac{N_{k*}+q_k}{N_{k*}})\Big]\sum_{j=1}^{q_m}\frac{1}{n}\Big[\prod_{1\le k<j}(1+\frac{1}{N_{m*}+q_m-k})\Big]\frac{N^t_{m*}}{N_{m*}+q_m-j} \\
&= [\prod_{1\le k<m}(\frac{N_{k*}+q_k}{N_{k*}})][\frac{q_m}{n}][\frac{N^t_{m*}}{(N_{m*})^{\beta_{m*}(t)}}] \quad \text{(see the third equality in (6.11))}.
\end{aligned}
$$

Then (6.12) and (6.13) yield (3.7) for m. This completes the proof of Lemma 6.1. □

Lemma 6.2. *Expression (3.16) in Theorem 3.1 equals* 0.

Proof: Denote by $\sum_{k=1}^m \Big\{s_k\Big\}$ expression (3.16). To prove the lemma, it suffices to show that each summand, s_k, in summation (3.16) vanishes.

It is trivial when $k = 1$. When $k \ge 2$, denoting $\sum_{j=1}^i a_j = 0$ if $i < 1$,

$$
s_k = \frac{\mathrm{P}\{X \in (l_k, t]\}\mathrm{P}\{(L,R) = (l_k, r_k)\}}{1 - \sum_{h<k}\mathrm{P}\{(L,R) = (l_h, r_h)\}} - \mathrm{P}\{X \in (l_k, t]\}\mathrm{P}\{(L,R) = (l_k, r_k)\}
$$

$$
\begin{aligned}
&-\sum_{j=1}^{k-1}\Big[\frac{\mathrm{P}\{(L,R)=(l_j,r_j)\}}{1-\sum_{h<j}\mathrm{P}\{(L,R)=(l_h,r_h)\}}\frac{\mathrm{P}\{X\in(l_k,t]\}\mathrm{P}\{(L,R)=(l_k,r_k)\}}{1-\sum_{i=1}^{j}\mathrm{P}\{(L,R)=(l_i,r_i)\}}\Big]\\
=&\frac{\mathrm{P}\{X\in(l_k,t]\}\mathrm{P}\{(L,R)=(l_k,r_k)\}}{1-\sum_{h<k}\mathrm{P}\{(L,R)=(l_h,r_h)\}}-\mathrm{P}\{X\in(l_k,t]\}\mathrm{P}\{(L,R)=(l_k,r_k)\}\\
&-\Big\{\frac{\mathrm{P}\{(L,R)=(l_{k-1},r_{k-1})\}}{1-\sum_{h<k-1}\mathrm{P}\{(L,R)=(l_h,r_h)\}}\frac{\mathrm{P}\{X\in(l_k,t]\}\mathrm{P}\{(L,R)=(l_k,r_k)\}}{1-\sum_{i=1}^{k-1}\mathrm{P}\{(L,R)=(l_i,r_i)\}}\Big\}\\
&-\sum_{j=1}^{k-2}\Big\{\frac{\mathrm{P}\{(L,R)=(l_j,r_j)\}}{1-\sum_{h<j}\mathrm{P}\{(L,R)=(l_h,r_h)\}}\frac{\mathrm{P}\{X\in(l_k,t]\}\mathrm{P}\{(L,R)=(l_k,r_k)\}}{1-\sum_{i=1}^{j}\mathrm{P}\{(L,R)=(l_i,r_i)\}}\Big\}\\
=&\Big\{\frac{1-\sum_{i=1}^{k-1}\mathrm{P}\{(L,R)=(l_i,r_i)\}}{1-\sum_{h<k-1}\mathrm{P}\{(L,R)=(l_h,r_h)\}}\frac{\mathrm{P}\{X\in(l_k,t]\}\mathrm{P}\{(L,R)=(l_k,r_k)\}}{1-\sum_{i=1}^{k-1}\mathrm{P}\{(L,R)=(l_i,r_i)\}}\Big\}\\
&-\sum_{j=1}^{k-2}\Big\{\frac{\mathrm{P}\{(L,R)=(l_j,r_j)\}}{1-\sum_{h<j}\mathrm{P}\{(L,R)=(l_h,r_h)\}}\frac{\mathrm{P}\{X\in(l_k,t]\}\mathrm{P}\{(L,R)=(l_k,r_k)\}}{1-\sum_{i=1}^{j}\mathrm{P}\{(L,R)=(l_i,r_i)\}}\Big\}\\
=&\frac{\mathrm{P}\{X\in(l_k,t]\}\mathrm{P}\{(L,R)=(l_k,r_k)\}}{1-\sum_{h<k-1}\mathrm{P}\{(L,R)=(l_h,r_h)\}}-\mathrm{P}\{X\in(l_k,t]\}\mathrm{P}\{(L,R)=(l_k,r_k)\}\\
&-\sum_{j=1}^{k-2}\Big\{\frac{\mathrm{P}\{(L,R)=(l_j,r_j)\}}{1-\sum_{h<j}\mathrm{P}\{(L,R)=(l_h,r_h)\}}\frac{\mathrm{P}\{X\in(l_k,t]\}\mathrm{P}\{(L,R)=(l_k,r_k)\}}{1-\sum_{i=1}^{j}\mathrm{P}\{(L,R)=(l_i,r_i)\}}\Big\}
\end{aligned}
$$

It is important to note that the last expression has the same pattern as the first expression, except that

$$
\sum_{j=1}^{k-1}\ \text{and}\ \sum_{h<k}\ \text{in the first one are replaced by}\ \sum_{j=1}^{k-2}\ \text{and}\ \sum_{h<k-1}\ \text{in the last one,}
$$

respectively. Thus inductively, we can show that $\sum_{j=1}^{k-1}$ and $\sum_{h<k}$ in the first expression can be replaced by $\sum_{j=1}^{1-1}$ and $\sum_{h<1}$. That is

$$
\begin{aligned}
s_k=&\frac{\mathrm{P}\{X\in(l_k,t]\}\mathrm{P}\{(L,R)=(l_k,r_k)\}}{1-\sum_{h<1}\mathrm{P}\{(L,R)=(l_h,r_h)\}}-\mathrm{P}\{X\in(l_k,t]\}\mathrm{P}\{(L,R)=(l_k,r_k)\}\\
&-\sum_{j=1}^{1-1}\Big\{\frac{\mathrm{P}\{(L,R)=(l_j,r_j)\}}{1-\sum_{h<j}\mathrm{P}\{(L,R)=(l_h,r_h)\}}\frac{\mathrm{P}\{X\in(l_k,t]\}\mathrm{P}\{(L,R)=(l_k,r_k)\}}{1-\sum_{i=1}^{j}\mathrm{P}\{(L,R)=(l_i,r_i)\}}\Big\}\\
=&\mathrm{P}\{X\in(l_k,t]\}\mathrm{P}\{(L,R)=(l_k,r_k)\}-\mathrm{P}\{X\in(l_k,t]\}\mathrm{P}\{(L,R)=(l_k,r_k)\}\\
=&0.\quad\square
\end{aligned}
$$

Lemma 6.3. *Statement (S3) in the proof of Theorem 4.1 holds.*

Proof: Denote $A_m^b=\mathcal{B}\setminus A_m$. We will show that

(S4) $\qquad A_m$ is a closed set,

thus, A_m^b is an open set, which equals $\cup_{i\geq 1}(a_i, b_i)$, where the (a_i, b_i)'s are disjoint open intervals and $a_i, b_i \in A_m \cup \mathcal{B}^c \cup \{\pm\infty\}$. Then we will show that

(S5) $$\mathrm{P}\{X \in (a_i, b_i)\} < 1/\sqrt{m} \text{ for any } i.$$

By the condition in the lemma, (4.2) and (4.3) hold for $t \in A_m$; moreover, (4.2) and (4.3) hold for $t \in \mathcal{B}^c$ by the arguments after inequality (4.3); furthermore, (4.2) and (4.3) hold if $t = \pm\infty$. The reason of the last statement is as follows. If $t = \pm\infty$, $\mathrm{P}\{L < t < R\} = 0$. Thus $N_k^t = 0$ and $Q_n(t) = 0$ (see (2.2) and (3.3)). It follows that $\mathrm{P}\{L < X \leq t < R\} = 0 = Q_n(t)$, which implies (4.2) and (4.3). As a consequence, if $t = a_i$ or b_i, (4.2) and (4.3) hold and

$$|\hat{S}_I(t) - S(t)| \leq O(1/\sqrt{m}) \quad \text{(see (3.3), (3.4) and (3.5))}.$$

Since both $\hat{S}_I(t)$ and $S(t)$ are nonincreasing function of t and in view of (S5), it is easy to show that $|\hat{S}_I(t) - S(t)| \leq O(1/\sqrt{m})$ for $t \in (a_i, b_i)$, which is equivalent to inequalities (4.2) and (4.3) for $t \in (a_i, b_i)$. Thus the proof of the lemma will be completed after we show (S4) and (S5).

To show (S4), let $\{t_k\}$ be a sequence of points in A_m which satisfies $\lim_{k\to\infty} t_k = t_0$. We need to show that $t_0 \in A_m$. By the notation in the proof of Theorem 4.1 (see the definition of A_m), for each t_k, $\mathrm{P}\{X \in [l_{m_k}, r_{m_k}]\} \geq \frac{4}{\sqrt{m}}$, where

$l_{m_k} = \sup\{l;\ l < t_k < r, (l, r) \in \mathcal{V}, \mathrm{P}\{l < X \leq t_0\} > 0\}$,
$r_{m_k} = \inf\{r;\ l < t_k < r, (l, r) \in \mathcal{V}, \mathrm{P}\{l < X \leq t_0\} > 0\}$.

It can be seen that the sequence of intervals (l_{m_k}, r_{m_k}) also satisfy Condition DI. Since the probability is bounded by 1, there are at most finitely many disjoint (l_{m_k}, r_{m_k}). WLOG, we can assume that $[l_{m_k}, r_{m_k}] \supset [l_{m_{k+1}}, r_{m_{k+1}}]$, $k \geq 1$. It follows that $\mathrm{P}\{X \in [l_{m_0}, r_{m_0}]\} \geq \frac{4}{\sqrt{m}}$, where $[l_{m_0}, r_{m_0}] = \cap_k [l_{m_k}, r_{m_k}]$ and it can be verified that

$l_{m_0} = \sup\{l;\ l < t_0 < r, (l, r) \in \mathcal{V}, \mathrm{P}\{l < X \leq t_0\} > 0\}$,
$r_{m_0} = \inf\{r;\ l < t_0 < r, (l, r) \in \mathcal{V}, \mathrm{P}\{l < X \leq t_0\} > 0\}$.

As a consequence, $t_0 \in A_m$. Thus, A_m is closed i.e., (S4) holds.

We now show (S5). Given (a_i, b_i), it follows from the definition of A_m^b that for each $t \in (a_i, b_i)$, there is an interval $(l_t, r_t) \in \mathcal{V}$ such that $\mathrm{P}\{X \in (l_t, r_t)\} < 4/\sqrt{m}$ and $l_t < t < r_t$. The collection $\{(l_t, r_t); t \in (a_i, b_i)\}$ is a cover of (a_i, b_i) i.e., $\cup_t (l_t, r_t) \supset (a_i, b_i)$. It follows that either (1) there is an $(l_t, r_t) \supset (a_i, b_i)$ or (2) there are three points $t_1, t_2, t_3 \in (a_i, b_i)$ and two intervals (l_{t_1}, r_{t_1}) and (l_{t_2}, r_{t_2}) in the collection $\{(l_t, r_t); t \in (a_i, b_i)\}$ such that $t_3 \in (l_{t_1}, r_{t_1}) \cap (l_{t_2}, r_{t_2})$, $t_1 \notin (l_{t_2}, r_{t_2})$, $t_2 \notin (l_{t_1}, r_{t_1})$ and $t_1 < t_3 < t_2$. However, case (2) is impossible since (l_{t_1}, r_{t_1}) and (l_{t_2}, r_{t_2}) violate Condition DI but they belong to $\mathcal{V}$, whose elements satisfy Condition DI. It follows that there exists at least an (l_t, r_t) in this collection such that $(l_t, r_t) \supset (a_i, b_i)$. Then $\mathrm{P}\{X \in (a_i, b_i)\} \leq \mathrm{P}\{X \in (l_t, r_t)\} < 4/\sqrt{m}$. Thus (S5) holds. □

Lemma 6.4 *Using the notation in the proof of Theorem 4.1, (4.6) and (4.7) hold.*

Proof: The proof of the lemma is an analog of the one for Lemma 6.1. In the following, in order to avoid exponents $\beta_{j*}(t)$ etc. (see (4.6)) we first assume that $N_{m_t*} \neq 0$. Thus $N_{j*} \neq 0$ for all $j \leq m_t$.

We first prove (4.6) and (4.7) for $m_t = 1$.

Let $Y_{j_1}^*, ..., Y_{j_{q_1^-}}^*$ be the Y_j^*'s which belong to D_1^-, then

$$N_{1*-} + q_1^- - k \leq N_{j_k} \leq N_{1*-} + q_1^- - k + \triangle_1, \tag{6.14}$$

$$\frac{N_{j_k}^t}{N_{j_k}} \le \frac{N_{1*}^t + \triangle_1}{N_{1*-} + q_1^- - k + \triangle_1},$$

$$\prod_{1\le j<k} \left(1 + \frac{1}{N_{1*-} + q_1^- - j + \triangle_1}\right) \frac{N_{1*-}^t}{N_{1*-} + q_1^- - k + \triangle_1}$$

$$\le \prod_{1\le j<j_k} \left(1 + \frac{1}{N_j}\right)^{\beta_j(t)} \frac{N_{j_k}^t}{N_{j_k}} \quad \left(\text{since } \left(1 + \frac{1}{N_j}\right)^{\beta_j(t)} = 1 \text{ if } \beta_j(t) = 0\right)$$

$$\le \prod_{1\le j<k} \left(1 + \frac{1}{N_{1*-} + q_1^- - j}\right) \frac{N_{1*-}^t + \triangle_1}{N_{1*-} + q_1^- - k + \triangle_1}, \tag{6.15}$$

for $k = 1, ..., q_1^-$. Let p_i be such that $L^*_{(p_i)} \le l_i < L^*_{(p_{i+1})}$. Summing up each term in (6.15) over $k = 1, ..., q_1^-$ yields

$$\sum_{k=1}^{q_1^-} \prod_{1\le j<k} \left(1 + \frac{1}{N_{1*-} + q_1^- - j + \triangle_1}\right) \frac{N_{1*-}^t}{N_{1*-} + q_1^- - k + \triangle_1}$$

$$\le \sum_{k=1}^{p_1 - q_1^+ - q_1^0} \prod_{1\le j<k} \left(1 + \frac{1}{N_j}\right)^{\beta_j(t)} \frac{N_k^t}{(N_k)^{\beta_k(t)}}$$

$$\le \sum_{k=1}^{q_1^-} \prod_{1\le j<k} \left(1 + \frac{1}{N_{1*-} + q_1^- - j}\right) \frac{N_{1*-}^t + \triangle_1}{N_{1*-} + q_1^- - k + \triangle_1}. \tag{6.16}$$

The second summation in (6.16) is due to $N_k^t = \beta_k(t) = 0$ if $k \le p_1 - q_1^+ - q_1^0$ and if $k \notin \{j_1, ..., j_{q_1^-}\}$. That is,

$$\sum_{k=1}^{q_1^-} \prod_{1\le j<j_k} \left(1 + \frac{1}{N_j}\right)^{\beta_j(t)} \frac{N_{j_k}^t}{N_{j_k}} = \sum_{k=1}^{p_1 - q_1^+ - q_1^0} \prod_{1\le j<k} \left(1 + \frac{1}{N_j}\right)^{\beta_j(t)} \frac{N_k^t}{(N_k)^{\beta_k(t)}}.$$

By an induction argument on q_1^0, we can show that

$$\sum_{k>p_1 - q_1^+ - q_1^0}^{p_1} \prod_{1\le j<k} \left(1 + \frac{1}{N_j}\right)^{\beta_j(t)} \frac{N_k^t}{N_k^{\beta_k(t)}} = \prod_{1\le j\le p_1 - q_1^+ - q_1^0} \left(1 + \frac{1}{N_j}\right)^{\beta_j(t)} q_1^0 \frac{N_{1*}^t}{N_{1*}}. \tag{6.17}$$

Simplifying the first and the third expression in (6.16) yields

$$q_1^- \frac{N_{1*-}^t}{N_{1*-} + \triangle_1} \le \sum_{k=1}^{p_1 - q_1^+ - q_1^0} \prod_{1\le j<k} \left(1 + \frac{1}{N_j}\right)^{\beta_j(t)} \frac{N_k^t}{(N_k)^{\beta_k(t)}} \le q_1^- \frac{N_{1*-}^t + \triangle_1}{N_{1*-} + \triangle_1}. \tag{6.18}$$

Inequality (6.14) also yields

$$\prod_{1\le j\le q_1^-}\Big(1+\frac{1}{N_{1*-}+q_1^- -j+\triangle_1}\Big)\le \prod_{1\le j\le p_1-q_1^+-q_1^0}\Big(1+\frac{1}{N_j}\Big)^{\beta_j(t)} \le \prod_{1\le j\le q_1^-}\Big(1+\frac{1}{N_{1*-}+q_1^- -j}\Big). \tag{6.19}$$

Simplifying the first and the third product in (6.19) yields

$$\Big(\frac{N_{1*-}+q_1^-+\triangle_1}{N_{1*-}+\triangle_1}\Big)\le \prod_{1\le j\le p_1-q_1^+-q_1^0}\Big(1+\frac{1}{N_j}\Big)^{\beta_j(t)}\le \Big(\frac{N_{1*-}+q_1^-}{N_{1*-}}\Big). \tag{6.20}$$

(6.17) and (6.20) yield

$$q_1^0\frac{N_{1*-}+q_1^-+\triangle_1}{N_{1*-}+\triangle_1}\frac{N_{1*}^t}{N_{1*}^{\beta_{1*}(t)}}\le \sum_{k>p_1-q_1^+-q_1^0}^{p_1}\prod_{1\le j<k}\Big(1+\frac{1}{N_j}\Big)^{\beta_j(t)}\frac{N_k^t}{N_k^{\beta_k(t)}} \le q_1^0\frac{N_{1*-}+q_1^-}{N_{1*-}}\frac{N_{1*}^t}{N_{1*}^{\beta_{1*}(t)}}. \tag{6.21}$$

(6.18) and (6.21) yield

$$q_1^-\frac{N_{1*-}^t}{N_{1*-}^{\beta_{1*-}(t)}+\triangle_1}+q_1^0\Big(\frac{N_{1*-}+q_1^-+\triangle_1}{N_{1*-}+\triangle_1}\Big)\frac{N_{1*}^t}{N_{1*}^{\beta_{1*}(t)}} \le \sum_{k=1}^{p_1}\prod_{1\le j<k}\Big(1+\frac{1}{N_j}\Big)^{\beta_j(t)}\frac{N_k^t}{(N_k)^{\beta_k(t)}} \le q_1^-\frac{N_{1*-}^t+\triangle_1}{N_{1*-}^{\beta_{1*-}(t)}+\triangle_1}+q_1^0\Big(\frac{N_{1*-}+q_1^-}{N_{1*-}}\Big)^{\beta_{1*}(t)}\frac{N_{1*}^t}{N_{1*}^{\beta_{1*}(t)}}. \tag{6.22}$$

It can be seen that if $N_{m_t*}=0$, (6.22) is still true. (6.22) is equivalent to (4.6) and (4.7) when $m_t=1$. (The above arguments are similar to the induction argument when $m=1$ in the proof of Lemma 6.1.)

Mimicing the arguments between (6.14) through (6.22), we can show inductively that (4.6) and (4.7) hold. □